# Learning multimodal latent attributes.

Fu, Y; Hospedales, TM; Xiang, T; Gong, S

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication.


# Learning Multi-modal Latent Attributes

Yanwei Fu, Timothy M. Hospedales, Tao Xiang and Shaogang Gong

**Abstract**—The rapid development of social media sharing has created a huge demand for automatic media classification and annotation techniques. Attribute learning has emerged as a promising paradigm for bridging the semantic gap and addressing data sparsity via transferring attribute knowledge in object recognition and relatively simple action classification. In this paper, we address the task of attribute learning for understanding multimedia data with sparse and incomplete labels. In particular we focus on videos of social group activities, which are particularly challenging and topical examples of this task because of their multi-modal content and complex and unstructured nature relative to the density of annotations. To solve this problem, we (1) introduce a concept of semi-latent attribute space, expressing user-defined and latent attributes in a unified framework, and (2) propose a novel scalable probabilistic topic model for learning multi-modal semi-latent attributes, which dramatically reduces requirements for an exhaustive accurate attribute ontology and expensive annotation effort. We show that our framework is able to exploit latent attributes to outperform contemporary approaches for addressing a variety of realistic multimedia sparse data learning tasks including: multi-task learning, learning with label noise, N-shot transfer learning and importantly zero-shot learning.

**Index Terms**—Attribute Learning, Latent Attribute Space, Multi-task Learning, Transfer Learning, Zero-shot Learning.

## 1 Introduction

With the rapid development of devices capable of digital media capture, vast volumes of multimedia data are uploaded and shared on social media platforms (e.g. YouTube and Flickr). For example, 48 hours of video are uploaded every minute in YouTube[1]. Managing this growing volume of data demands effective techniques for automatic media understanding. Such automatic techniques are important for content based understanding in order to enable effective indexing, search, retrieval, filtering and recommendation of multimedia content from the vast quantity of social images and video data.

Content based understanding aims to model and predict classes and tags relevant to objects, sounds and events – anything likely to be used by humans to describe or search for media. One of the most common but most challenging types of data for content analysis is that of unstructured social group activity, which is common in consumer video (e.g. home videos) [18]. The unconstrained space of objects, events and interactions in consumer videos makes them intrinsically more complex than commercial videos (e.g. movies, news and sports). This unconstrained domain gives rise to a space of possible content concepts that is orders of magnitude greater than that typically addressed by most previous video analysis work (e.g. human action recognition). Furthermore, the casual nature of consumer videos makes it difficult to extract good features: they are typically captured with low resolution, poor lighting, occlusion, clutter, camera shake and background noise.

• *The authors are with the School of Electronic Engineering and Computer Science, Queen Mary University of London, E1 4NS, UK.*
*Email: {yanwei.fu,tmh,txiang,sgg}@eecs.qmul.ac.uk*

1. http://www.youtube.com/t/faq

To tackle these problems, we wish to learn a model capable of content based prediction of class and tag annotations from multi-modal video data. However the ability to learn good annotation models is often limited in practice by insufficient and poor quality training annotations. The underlying challenges here can be broadly characterised as *sparsity*, *incompleteness* and *ambiguity*.

**Annotations are sparse.** Consumer media covers a huge unconstrained space of object/activity/event concepts, therefore requiring numerous tags to completely annotate the underlying content. However the number of labelled training instances per annotation concept is likely to be low. For example, consumer videos shared on social media platforms only have 2.5 tags on average versus 9 tags in general YouTube videos [18].

**Annotations are intrinsically incomplete.** Since the space of concepts is unconstrained, exhaustive manual annotation of examples for every concept is impractically expensive, even through mechanisms such as Amazon Mechanical Turk (AMT) [35]. Previous studies have therefore focused on analyzing relatively constrained spaces of content and hence annotation ontologies [24]. However, there are for example, some 30000 relevant object classes which are recognizable by humans [3]. This means that any ontology will either be too small to provide a complete vocabulary to describe general videos, or have insufficient training data for every concept.

**Annotations are ambiguous.** *Ambiguity* is relatively less studied in previous work but a significant challenge for semantic media understanding. Even for the same image/video, subjective factors (e.g. cultural background) may lead to contradictory and ambiguous annotations. A well-known example is that some countries take nodding head as "yes", while others as "no". This ambiguity of annotations can be taken as label noise. Ambiguity also arises from the semantic gap between

Digital Object Indentifier 10.1109/TPAMI.2013.128 

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication.


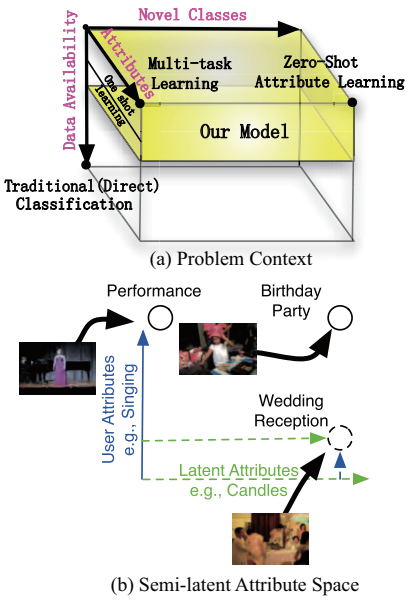


Figure 1. (a) Learning a semi-latent attribute space is applicable to various problem domains. (b) Representing data in terms of a semi-latent attribute space partially defined by the user (solid axes), and partially learned by the model (dashed axes). A novel class (dashed circle) may be defined in terms of both user-defined and latent attributes.

annotations and raw data: semantically obvious annotations are not necessarily detectable from low-level features; while the most useful annotations for a model may not be the most semantically obvious ones that humans commonly provide. Finally the weakly supervised nature of annotation, and the multi-modality of the data are another strong sources of ambiguity, e.g., an annotation of "clapping" comes with no information detailing where it was observed (temporally) in a video, or whether it was only seen visually, only heard, or both seen and heard.

One strategy to address the sparsity of annotation is via exploitation of shared components or parts between different classes. For example, in an object recognition context a wheel may be shared between a car and a bicycle [9]; while in an activity context, "bride" can be seen in classes of "wedding ceremony", "wedding dance" and "wedding reception". These shared parts are often referred to as attributes. Attributes focus on *describing* an instance (e.g., has legs) rather than *naming* it (e.g., is a dog), and they provide a semantically meaningful bridge between raw data and higher level classes. The concept of attributes can be traced back to the early work of intrinsic images [2], but attribute learning has been popularized recently as a powerful approach for image and video understanding with sparse training examples [22], [10], [9], [31], [30]. Most previous work has looked at attributes as a solution to *sparseness* of annotation, but focused on constrained domains and single modalities, avoiding the bigger issues in intrinsic *incompleteness* and *ambiguity*. This paper shows that attributes not only can help to solve sparsity, but also assist in overcoming the intrinsic incompleteness and ambiguity of annotation.

To address these challenges, we introduce a new attribute learning framework (Fig. 1) which learns a unified *semi-latent attribute space* (Fig. 1(b)). *Latent attributes* represent all shared aspects of the data which are not explicitly included in users' sparse and incomplete annotations. These are complementary to user-defined attributes, and discovered automatically by a model through jointly learning of the semi-latent attribute space (see Section 4.2). This learned space provides a mechanism for *semantic feature reduction* [30] from the raw data in multiple modalities to a unified lower dimensional semantic attribute space (Fig. 1(b)). The semi-latent space bridges the semantic gap with reduced dependence on the completeness of the attribute ontology and accuracy of the training attribute labels. Fig. 1(a) highlights this property by putting our work in context of various standard problems. Our semi-latent attribute space consists of three types of attributes: user-defined (UD) attributes from any prior concept ontology; latent class-conditional (CC) attributes [15] which are discriminative for known classes; and latent generalized free (GF) attributes [13] which represent shared aspects not in the attribute ontology. Jointly learning this unified space is important to ensure that latent CC and GF attributes represent unmodeled aspects of the data rather than merely rediscovering user-defined attributes.

To learn the semi-latent attribute space, we propose a multi-modal latent attribute topic model (M2LATM), building on probabilistic topic models [7], [15]. M2LATM jointly learns user-defined and latent attributes, providing an intuitive mechanism for bridging the semantic gap and modeling sparse, incomplete and ambiguous labels. To learn the three types of attributes, the model learns three corresponding sets of topics with different constraints. UD topics are constrained in 1 to 1 correspondence with attributes from the ontology. Latent CC topics are constrained to match the class label while latent GF topics are unconstrained. Multi-task classification, N-shot learning and zero-shot learning are performed in the learned semantic attribute space. To make the learning and inference scalable, we exploit equivalence classes for scalability by expressing our topic model in "vocabulary" rather than "word" domain.

## 2 Related Work

**Semantic concept detection** Studies addressing concept detection [34], [12] (also known as tagging [13], [38], [41], and multi-label classification [32], image [39] and video [32], [36] annotation) are related to attribute-learning [22], [24]. Concept detection has been quite extensively studied, and there are standard benchmarks such as

TRECVID[2], LSCOM[3] and MediaMill[4]. One way to contrast these bodies of work is that these studies typically predict tags for the purpose of indexing for content based search and retrieval. In contrast, attribute-learning studies typically focus on how learned attributes can be re-used or transferred to other tasks or classes. Depending on the ontology, level of abstraction and model used, many annotation approaches can therefore be seen as addressing a sub-task of attribute-learning. Some annotation studies aim to automatically expand [12] or enrich [41] the set of tags queried in a given search. This is a related motivation to our latent attributes. However, the possible space of expanded/enriched tags is still constrained by fixed ontology and may be very large (e.g., vocabulary space of over $20,000$ tags in [38]), which are constraints we aim to relax.

Recently, mid-level semantic concept detectors based on video ontologies have also been used to provide additional cues for high-level event detection. For example, various submissions [19], [8] to the TRECVID Multimedia Event Detection (MED)[5] challenge have successfully exploited variants of this strategy. In this context, semantic concept detectors are related to the idea of user-defined attributes. However, these studies generally consider huge and strongly-labelled datasets with exhaustive and prescriptive ontologies; whereas we aim to learn from sparse data with incomplete ontologies.

**Attribute Learning** A key contribution of attribute-based representations has been to provide an intuitive mechanism for multi-task [33] and transfer [16] learning: enabling learning with few or zero instances of each class via sharing attributes. Attribute-centric semantic models of data have been explored for images [22], [10] and to a lesser extent video [24]. Applications include modeling properties of human actions [24], animals [22], faces [21], scenes [16], and objects [9], [10]. However, most of these studies [22], [9], [16], [30], [26], [20] assume that an exhaustive space of attributes has been manually specified. In practice, an exhaustive space of attributes is unlikely to be available, due to the expense of ontology creation, and that semantically obvious attributes for humans do not necessarily correspond to the space of detectable and discriminative attributes [31] (Fig. 1(b)). One method of collecting labels for large scale problems is to use AMT [35]. However, even with excellent quality assurance, the results collected still exhibit strong label noise. Thus label-noise [37] is a serious issue in learning from either AMT, or existing social meta-data. More subtly, even with an exhaustive ontology, only a subset of concepts from the ontology are likely to have sufficient annotated training examples, so the portion of the ontology which is effectively usable for learning may be much smaller.

Fig. 2 contrasts Direct attribute prediction (DAP [22])

2. http://www-nlpir.nist.gov/projects/trecvid/
3. http://www.lscom.org/
4. http://www.science.uva.nl/research/mediamill/challenge/
5. http://www.nist.gov/itl/iad/mig/med12.cfm

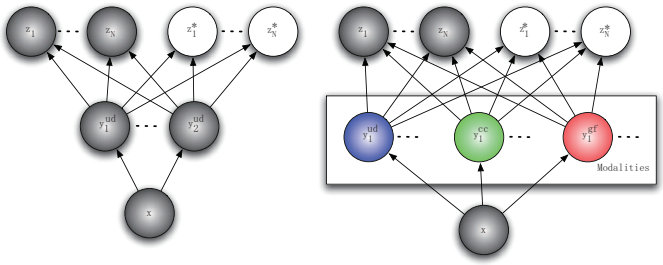


Figure 2. Schematic of conventional (left) DAP [22] versus (right) M2LATM. Shading indicates different types of constraints placed on the variables. Symbols are explained in Section 4.

– a popular attribute learning framework – with our M2LATM. The shading indicates the types of constraints placed on the nodes, with the dark nodes being fully observed, and the colored nodes in M2LATM having UD, CC and GF type constraints. A few studies [10], [24] augmented user-defined (UD) attributes by data-driven attributes, similar to CC attributes, to better differentiate existing classes. However, our more nuanced distinction between CC and GF latent attributes better helps differentiate both existing classes and novel classes: CC are limited to those which differentiate existing classes; without this constraint, GF attributes provide an additional cue to help differentiate novel classes. Previous work [10], [24] also learns UD and CC attributes separately. This means that the learned CC attributes are not necessarily complementary to the user-defined ones (i.e., they may be redundant). Finally, we also uniquely show how to use latent attributes in zero-shot learning.

To the best of our knowledge, prior work has focused on single modalities, e.g. static appearance. Building attribute models of multimedia video requires special care to ensure all content modalities (such as static appearance (e.g. salient objects), motion (e.g. human actions) and auditory (e.g. songs)) are coherently and fully exploited. A powerful class of models for generatively modelling multiple modalities of data and low-dimensional representations such as attributes is that of topic modelling, which we discuss next.

**Topic Models** Probabilistic topic models (PTMs) [7] have been used extensively in modeling images [39] and video [40], [29] via learning a low-dimensional topic representation. PTMs are related to attribute learning in that multiple tags can be modeled generatively [4], [39], and classes can be defined in terms of their typical topics [39], [6], [15], [13]. However these topic-representations are generally discovered automatically and lack the semantic meaning which attribute models obtain by supervising the intermediate representation. There has been limited work [43], [11] using topics to directly represent attributes, and provide attractive properties of attribute learning such as zero-shot learning. These are limited to user-defined attributes only [43], or formulated in a computationally non-scalable way and for a single modality only [43], [11]. In contrast to [43] (as well as

most annotation studies [37], [36], [34], [32]), we leverage the ability of topic models to learn unsupervised representations from data; and in contrast to [40], [39], [29], [6], our framework also leverages prior knowledge of user-defined classes and attributes. Together, these properties provide a complete and powerful semi-latent semantic attribute-space. Scalability can also be a serious issue for topic models applied to video, as most formulations take time proportional to the volume of features [7], [39], [43], [11]. Our unstructured social activity attribute (USAA) dataset [11] is bigger than huge text datasets which have been addressed with large-scale distributed algorithms and supercomputers [28]. We therefore generalize ideas in sparse equivalence class updating [1] to make inference tractable in M2LATM.

### 2.1 Contributions

By extending our preliminary work reported in [11], this paper formulates systematically a semi-latent attribute space learning framework that makes the following specific contributions: (i) We address a key problem in attribute learning from sparse, incomplete and ambiguous annotation – focusing on multi-modal social group activities captured in unstructured and complex consumer videos, notably different from previously reported work. (ii) We introduce a semi-latent attribute space, which enables the use of as much or as little prior knowledge as available from both user-defined and the two types of automatically discovered latent-attributes. (iii) We formulate a computationally tractable solution of this strategy via a novel and scalable topic model. (iv) We show how latent attributes computed by our framework can be utilised to tackle a wide variety of learning tasks in the context of multimedia content understanding including multi-task, label-noise, N-shot and surprisingly zero-shot learning. (v) We provide extensive evaluation of the proposed model against contemporary methods for a variety of challenging datasets.

## 3 Video Feature Extraction and Representation

The foundation for video content understanding is extracting and representing suitably informative and robust features. This is especially challenging for unconstrained consumer video and unstructured social activity due to dramatic within-class variations, as well as noise sources of occlusion, clutter, poor lighting, camera shake and background noise [17]. Global features provide limited invariance to these noise-sources. Local keypoint features collected into a bag-of-words (BoW) are considered state of the art [18], [17], [41]. We follow [18], [17], [41], in extracting features for three modalities, namely static appearance, motion, and auditory. Specifically, we employ scale-invariant feature transform (SIFT) [25], spatial-temporal interest points (STIP) [23], and mel-frequency cepstrum (MFCC) respectively[6].

6. Refer to [18], [17], [41] for full feature extraction details.

## 4 Methods

### 4.1 Problem Context and Definition

We first formally introduce the problem of attribute-based learning before developing our contributions in the next section. Learning to detect or classify can be formalised as learning a mapping $F : \mathcal{X}^d \rightarrow \mathcal{Z}$ of $d$-dimensional raw data $\mathcal{X}$ to label $\mathcal{Z}$ from training data $D = \{(\mathbf{x}_i, z_i)\}_{i=1}^n$. A variant of the standard approach considers a composition of two mappings [30]:

$$F \;=\; S(L(\cdot)),\; L : \mathcal{X}^d \rightarrow \mathcal{Y}^p,\; S : \mathcal{Y}^p \rightarrow \mathcal{Z}, \qquad (1)$$

where $L$ maps the raw data to an intermediate representation $\mathcal{Y}^p$ (typically with $p \ll d$) and then $S$ maps the intermediate representation to the final class $\mathcal{Z}$. Examples of this approach include dimensionality-reduction via PCA (where $L$ is chosen to explain the variance of $\mathbf{x}$ and $\mathcal{Y}^p$ is the space of orthogonal principal components of $\mathbf{x}$) or linear discriminants and multi-layer neural networks (where $L$ is optimised to predict $\mathcal{Z}$).

Attribute learning [22], [30] exploits the idea of requiring $\mathcal{Y}^p$ to be a *semantic attribute* space. $L$ and $S$ are then learned by direct supervision with instance, attribute vector and class tuples $D = \{(\mathbf{x}_i, \mathbf{y}_i, z_i)_{i=1}^n\}$. This has benefits for sparse data learning including multi-task, N-shot and zero-shot (Fig. 1(a)). In multi-task learning [33] the statistical strength of the whole dataset can be shared to learn $L$, even if only subsets corresponding to particular classes can be used to learn each class in $S$. In N-shot transfer learning, the mapping $L$ is first learned on a large "source/auxiliary" dataset $D$. We can then effectively learn a much smaller "target" dataset $D^* = \{(\mathbf{x}_i, z_i^*)\}_{i=1}^m$, $m \ll n$ containing novel classes $z^*$ by transferring the attribute mapping $L$ to the target task, leaving only parameters of $S$ to be learned from the new dataset $D^*$. The key unique feature of attribute learning is that it allows zero-shot learning: the recognition of novel classes without any training examples $F : \mathcal{X}^d \rightarrow \mathcal{Z}^*$ ($\mathcal{Z}^* \notin \mathcal{Z}$) via the learned attribute mapping $L$ and a manually specified attribute description $S^*$ of the novel class.

### 4.2 Semi-latent Semantic Attribute Space

Most prior attribute learning work (Sections 2 and 4.1), unrealistically assumes that the attribute space $\mathcal{Y}^p$ is completely defined in advance, and contains sufficiently many attributes which are both *reliably detectable* from $\mathcal{X}$ and *discriminative* for $\mathcal{Z}$. We now relax these assumptions by performing *semantic feature reduction* [30] from the raw data to a lower dimensional *semi-latent semantic attribute space* (illustrated in Fig. 1(b)).

**Definition 1.** Semi-latent semantic attribute space

*A $p$ dimensional metric space where $p_{ud}$ dimensions encode manually specified semantic properties, and $p_{la}$ dimensions encode latent semantic properties determined by some objective given the manually defined dimensions.*

We aim to define an attribute-learning model $L$ which can learn a semi-latent attribute space from training data $D$ where $|\mathbf{y}| = p_{ud}$, $0 \leq p_{ud} \leq p$. That is, only a $p_{ud}$ sized subset of the attribute dimensions are user-defined, and $p_{la}$ other relevant latent dimensions are discovered automatically. The attribute-space is thus partitioned into observed and latent subspaces: $\mathcal{Y}^p = \mathcal{Y}_{ud}^{p_{ud}} \times \mathcal{Y}_{la}^{p_{la}}$ with $p = p_{ud} + p_{la}$. To support a full spectrum of applications, the model should allow: (1) an exhaustively and correctly specified attribute space $p = p_{ud}$ (corresponding to previous attribute learning work); (2) a partially known attribute space $p = p_{ud} + p_{la}$ (corresponding to an incomplete ontology); and (3) a completely unknown attribute space $p = p_{la}$. Such a model would go beyond existing approaches to bridge the gap (Fig. 1(a)) between exhaustive and unspecified attribute ontologies. As we will see, performing classification in this semi-latent space will provide increased robustness to the amount of domain-knowledge/ontology creation budget, and to annotation noise as compared to conventional approaches.

### 4.3 Multi-modal Latent Attribute Topic model

To learn a suitable attribute model $L$ (Eq. (1)) with the flexible properties outlined in the previous section, we will build on probabilistic topic models [7], [15]. Essentially we will represent each attribute with one or more topics, and add different types of constraints to the topics such that some topics will represent user-defined attributes, and others latent attributes.

First, we briefly review the standard Latent Dirichlet Allocation (LDA) [7] approach to topic modeling. Applied to video understanding [14], [15], [13], [29], conventional LDA learns a generative model of videos $\mathbf{x}_i$. Each quantized feature $x_{ij}$ in clip $i$ is distributed according to a discrete distribution $p(x_{ij}|\boldsymbol{\beta}_{y_{ij}}, y_{ij})$ with a Dirichlet parameter $\boldsymbol{\beta}$ corresponding to its (unknown) parent topic $y_{ij}$. Topics in video $i$ are distributed according to another discrete distribution $p(\mathbf{y}_i|\boldsymbol{\theta}_i)$ parameterized by the Dirichlet variable $\boldsymbol{\theta}_i$. Finally, the prior probability of topics in a video are distributed according to the $p(\boldsymbol{\theta}_i|\alpha)$ with parameter $\alpha$.

Standard LDA is uni-modal and unsupervised. Unsupervised LDA topics can potentially *represent* fully latent (GF) attributes. We will modify LDA to constrain a subset of topics (UD and CC) to *represent* conventional supervised attributes [22], [30]. The three attribute types are thus given a concrete representation in practice by a single topic model with three types of topics (UD, CC and GF), differing in terms of the constraints with which they are learned. We next detail our M2LATM including learning from (1) supervised attribute annotations and (2) multiple modalities of observation.

#### 4.3.1 Attribute-topic model

In order to model supervised user-defined attribute annotations, M2LATM establishes a topic-attribute correspondence so that attribute $k$ is represented by topic $k$. We encode the (user-defined) attribute annotation for video $i$ via a per-instance vector topic prior $\boldsymbol{\alpha}_i$. An attribute $k$ is encoded as absent via setting $\alpha_{ik} = 0$, or present via $\alpha_{ik} = 1$. The full joint distribution for a database $D$ of videos with attribute annotations $\boldsymbol{\alpha}_i$ is:

$$p(D|\{\boldsymbol{\alpha}\}, \boldsymbol{\beta}) = \prod_i \int \left( \prod_j \sum_{y_{ij}} p(x_{ij}|y_{ij}, \boldsymbol{\beta}) p(y_{ij}|\boldsymbol{\theta}_i) \right) p(\boldsymbol{\theta}_i|\boldsymbol{\alpha}_i) d\theta_i, \quad (2)$$

To infer the attributes for a clip, we require the posterior $p(\boldsymbol{\theta}_i, \mathbf{y}_i|\mathbf{x}_i, \boldsymbol{\alpha}_i, \boldsymbol{\beta})$. As for LDA [7], this is intractable to compute exactly. Variational inference approximates the full posterior with a factored variational distribution:

$$q(\boldsymbol{\theta}_i, \mathbf{y}_i|\boldsymbol{\gamma}_i, \boldsymbol{\phi}_i) = q(\boldsymbol{\theta}_i|\boldsymbol{\gamma}_i) \prod_j q(y_{ij}|\phi_{ij}). \quad (3)$$

where $\gamma_{ik}$ parameterizes the Dirichlet factor of topic/attribute $k$ proportions $\boldsymbol{\theta}_i$ within clip $i$; and $\phi_{ijk}$ parameterizes the discrete posterior $y_{ij}$ of topic/attributes for feature $x_{ij}$. Optimizing the variational bound results in the updates:

$$\begin{aligned} \phi_{ijk} &\propto \beta_{x_{ij}k} \exp(\Psi(\gamma_{ik})), \\ \gamma_{ik} &= \alpha_{ik} + \sum_j \phi_{ijk}, \end{aligned} \quad (4)$$

where $\Psi$ is the digamma function. Iterating Eq. (4) to convergence completes the variational E-step of an expectation maximisation (EM) algorithm. The M-step updates parameter $\beta$ by maximum likelihood: $\beta_{vk} \propto \sum_{i,j} \mathbf{I}(x_{ij} = v)\phi_{ijk}$. After EM learning, each attribute/topic $y$ (e.g., clapping hands or singing) will be associated with a particular subset of the low-level features via $p(x|y, \boldsymbol{\beta})$ and learned parameter $\boldsymbol{\beta}$.

#### 4.3.2 Learning multiple modalities

Topic model generalizations exist to jointly model multiple translations of the same text [27] via a common topic profile $\boldsymbol{\theta}$, where one language could be considered one modality. However, this is insufficient because as we have discussed, a given attribute may be unique to a particular modality. To model multi-modal data $D = \{D_m\}_{m=1}^M, D_m = \{\mathbf{x}_{im}\}$, we therefore exploit a unique topic prior $\boldsymbol{\theta}_m$ per-modality $m$ as follows:

$$\begin{aligned} p(\{D_m\}|\{\boldsymbol{\alpha}\}, \{\boldsymbol{\beta}_m\}) &= \prod_{i,m} \int d\boldsymbol{\theta}_{im} p(\boldsymbol{\theta}_{im}|\boldsymbol{\alpha}_i) \\ &\times \left( \prod_j \sum_{y_{ijm}} p(x_{ijm}|y_{ijm}, \boldsymbol{\beta}_m) p(y_{ijm}|\boldsymbol{\theta}_{im}) \right). \end{aligned} \quad (5)$$

By sharing the annotations $\boldsymbol{\alpha}$ across modalities, but allowing a unique per-modality prior $\boldsymbol{\theta}_m$, the model is able to represent both attributes with strong multi-modal correlates (e.g., clapping hands) and those more unique to a particular modality (e.g., laughter, candles).

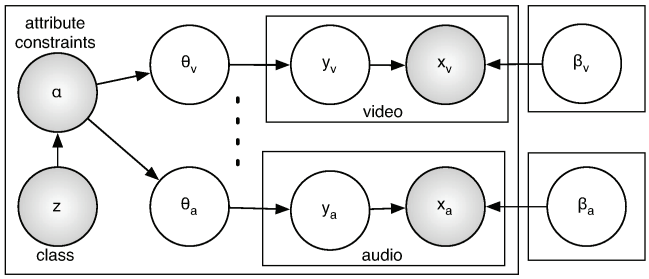


Figure 3. Graphical model for M2LATM.

Moreover, this approach provides an automatic way to deal with different modalities being expressed on different scales. Different scale modalities[7] is a serious problem for most topic models hoping to simply concatenate multi-modal data: either one modality dominates or words underflow is risked if data is normalized. For this reason studies [43] often only use a single modality when many are available. Fig. 3 provides a graphical model representation of M2LATM.

#### 4.3.3 Learning user-defined and latent attributes

With no user-defined attributes ($p = p_{la}, p_{ud} = 0$), a $p$-topic LDA model provides a mapping $L$ from raw data $\mathbf{x}$ to a $p$-dimensional latent space by way of the variational posterior $q(\boldsymbol{\theta}|\boldsymbol{\gamma})$. This is a discrete analogy to the common use of PCA to reduce the dimension of continuous data. However, to (i) support user-defined attributes when available and (ii) ensure the latent representation is discriminative, we add constraints.

User-defined attributes are typically provided in terms of length $p^{ud}$ binary vectors $\mathbf{v}^{ud}$ specifying the attributes of class $z$ or instance $i$ [22], [30]. We have no prior knowledge of the relation between $\mathbf{v}^{ud}$ and each word $(i, j)$, so $\mathbf{v}^{ud}$ cannot determine $\mathbf{y}$ directly. To enforce the user-defined attribute constraint, we define a *per instance* prior $\boldsymbol{\alpha}_i = [\boldsymbol{\alpha}_i^{ud}, \boldsymbol{\alpha}_i^{la}]$, setting $\alpha_{i,k}^{ud} = 0$ if $v_{i,k}^{ud} = 0$ and $\alpha_{i,k}^{ud} = 1$ otherwise. It enforces that instances $i$ lacking an attribute $k$ can never use that attribute to explain the data; but otherwise leaving the model to infer attribute proportions, modality and word correspondence.

To learn the latent portion of the attribute-space, we could simply leave the remaining portion $\boldsymbol{\alpha}^{la}$ of the prior unconstrained. However, for the latent space to be useful, it should be both *discriminative* (for class) and *generalizable* (to potential new classes) [15], [13]. To obtain both of these properties, we split the prior into components for "class-conditional" (CC) and "generalized free" (GF) topics. When learned jointly with UD attributes and with appropriate constraints, CC topics will be selective for known classes and GF topics will represent attributes shared between known classes, and hence likely to generalize. Specifically, we split the latent space prior $\boldsymbol{\alpha}_i^{la} = [\boldsymbol{\alpha}_i^{cc}, \boldsymbol{\alpha}^{gf}]$. In the CC component $\boldsymbol{\alpha}_i^{cc} = \{\boldsymbol{\alpha}_{i,z}\}_{z=1}^{N_z}$, each subset $\boldsymbol{\alpha}_{i,z}$ corresponds to a class $z$. For an instance $i$ with label $z_i$, set $\alpha_{i,z=z_i}^{cc} = 1$ and all other $\alpha_{i,z\neq z_i}^{cc} = 0$. This enforces that only instances with class $z$ can allocate topics $y_z^{cc}$ and hence that these topics are discriminative for class $z$. The GF component of the latent space prior is uniform $\boldsymbol{\alpha}^{gf} = 1$, meaning that GF topics are shared between all classes and thus represent aspects shared among all the data.

7. For example, 99% of the feature frequencies in USAA are in the range of $[0, 8]$ (appearance),$[0, 450]$ (motion), and$[0, 50]$ (auditory).

#### 4.3.4 Classification

To use M2LATM for classification, we define the mapping $L$ in Eq. (2) as the posterior statistic $\boldsymbol{\gamma}$ in Eq. (9). The remaining component to define is the attribute-class mapping $S$. Importantly, for our complex data, this mapping is not deterministic (i.e., 1:1 correspondence between attributes and classes assumed in [22], [30]). Like [24], we therefore use standard classifiers to learn this mapping from $(z_i, \boldsymbol{\gamma}_i)$ pairs obtained from our M2LATM attribute learner.

#### 4.3.5 Surprising attributes

M2LATM can also be used to find videos which exhibit surprising/abnormal semantics. Given the training labels and estimated set of posterior semi-latent topic profiles $\{z_i, \boldsymbol{\gamma}_i\}$, we can fit a multi-variate Gaussian $\mathcal{N}(\boldsymbol{\mu}_\gamma^z, \Sigma_\gamma^z)$ to the profile of examples from each class $z$. At test time, once the class $z^*$ of a given instance is estimated, we can detect surprising attribute semantics by computing the likelihood $p(\boldsymbol{\gamma}^*|\boldsymbol{\mu}_\gamma^{z^*}, \Sigma_\gamma^{z^*})$. Importantly, unlike earlier notions of attribute-surprise [10], this approach (i) also considers surprising latent attributes, and (ii) inter-attribute and inter-modality correlations.

### 4.4 Semi-latent Zero Shot Learning

Zero-shot learning addresses classification of unseen classes via semantic attribute descriptions rather than via learning from training examples. A description $\mathbf{v}_{z^*}^{ud} \in \mathcal{Y}_{ud}$ for a new class $z^*$ is provided in terms of attributes from human prior knowledge. Existing approaches [22], [30] define simple deterministic prototypes $\mathbf{v}_{z^*}^{ud}$ in terms of UD attributes only, and classify by matching these templates $\mathbf{v}_{z^*}^{ud}$ to the estimated UD attributes for each test instance, e.g., by nearest-neighbour (NN) [10] or naive-Bayes. Using NN, conventional zero-shot classification of test instance $\mathbf{x}^*$ with UD attribute representation $\mathbf{y}^{*,ud}$ is:

$$f(\mathbf{x}^*) \quad = \quad \arg\min_{z^*} \left\|\mathbf{y}^{*,ud} - \mathbf{v}_{z^*}^{ud}\right\|. \qquad (6)$$

However, in this approach one needs a large ontology of attributes, and to specify an (impractically long) definition of each new class in terms of every attribute in the ontology. Counter-intuitively, we can work with a smaller UD ontology $(p_{ud} \geq 1)$ and leverage the latent portion of the attribute-space to still obtain a rich representation for classification. We project a short/incomplete UD attribute description of a novel

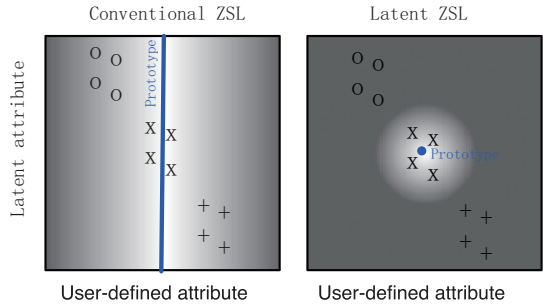


Figure 4. Schematic illustration of latent ZSL mechanism

class into the complete semi-latent attribute space description as follows:

1) Input a test set $D^* = \{\mathbf{x}^*\}$ containing novel classes, and UD attribute prototypes $\mathbf{v}_{z^*}^{ud}$ for those classes.
2) Infer attributes $\mathbf{y}^* = [\mathbf{y}^{*,ud}, \mathbf{y}^{*,la}]$ for each test data $\mathbf{x}^*$ (given by $\gamma$ in Eq. (4))
3) Let $NN_k^{ud}(\mathbf{v}_{z^*}^{ud}, \{\mathbf{y}^{*,ud}\})$ denote the set of $k$ nearest UD neighbours in $D^*$ to each prototype $\mathbf{v}_{z^*}^{ud}$ in $\mathcal{Y}_{ud}$.
4) Project UD prototypes $\mathbf{v}_{z^*}^{ud} \in \mathcal{Y}_{ud}$ into the full attribute space $\mathcal{Y}$ by averaging their nearest neighbours (Eq. (7)).
5) Perform zero-shot classification in the full attribute space $\mathcal{Y}$ (Eq. (8)).

$$\mathbf{v}_{z^*} = \frac{1}{k} \sum_{\mathbf{y} \in NN_k^{ud}(\mathbf{v}_{z^*}^{ud}, D^*)} \mathbf{y}, \quad (7)$$

$$f(\mathbf{x}^*) = \arg\min_{z^*} \|\mathbf{y}^* - \mathbf{v}_{z^*}\|. \quad (8)$$

The mechanism of this algorithm is schematically illustrated in two dimensions by Fig. 4. The one dimensional UD prototype $y^{*,ud}$ (blue line) only weakly identifies (shading) the target class 'x'. After projecting into the full space, the two-dimensional prototype (blue dot) more clearly identifies (shading) the target class.

Our approach can be viewed in a few ways: as transductively exploiting the test data distribution; or as one iteration of an EM-style algorithm for data with partially-known parameters and unknown variables (in contrast to the typical semi-supervised learning case of partially known variables and unknown parameters [44]). Previous ZSL studies are constrained to user-defined attributes, thus being critically dependent on the completeness of the user attribute-space. In contrast, our approach uniquely leverages a potentially much larger body of latent attributes via a loose manual definition of a novel class. We will show later this approach can significantly improve zero-shot learning performance.

### 4.5 Efficient Inference and Implementation

Our formulation thus far, as well as the earlier work [11] and LDA in general[8], infers the posterior over topics/attributes *for each word* (i.e. Eq. (4) indexed by word $j$). Computation is thus $\mathcal{O}(NK)$ for $N$ total words and $K$ topics. For our video dataset, where words correspond to dense interest point detections, $N$ is of the order $10^{10}$ and grows with video length. Conventional topic models do not scale to this data in either processing or memory demands, requiring days to run on in practice. In contrast, approaches such as support vector machines (SVM) [18] use the same data, but operate on word proportions within the vocabulary $V$. SVMs are thus $\mathcal{O}(V)$ and therefore significantly faster than conventional $\mathcal{O}(N)$ topic models because typically $V$ is $\leq 10^4$.

Inspired by [1], we observe that while each observation $x_{ijm}$ has an associated topic posterior, all instances of the same vocabulary item $x \in V$ within one video have the same posterior $\phi$. Exploiting this equivalence class, the same inference can therefore be expressed in the $\mathcal{O}(V)$ vocabulary domain, rather than the $\mathcal{O}(N)$ word domain. Inference for multiple modalities $m$ expressed in vocabulary-domain is thus:

$$\begin{aligned} \phi_{ivkm} &\propto \beta_{vkm} \exp(\Psi(\gamma_{ikm})), \\ \gamma_{ikm} &= \alpha_{ik} + \sum_v \mathbf{h}_v(\mathbf{x}_{im}) \phi_{ivkm}. \end{aligned} \quad (9)$$

Here, $\mathbf{h}_v(\mathbf{x}_{im})$ denotes the histogram of observations in $\mathbf{x}_{im}$, and the topic posterior matrix $\boldsymbol{\phi}_{x \cdot\cdot m}$ is now of size $VK$ instead of $NK$. Further efficiencies may be obtained by observing that only sufficient statistics for vocabulary elements observed in each document need to be computed. That is, Eq. (9) can be updated as a sparse matrix operation for unique observations $U_i$ at $\mathcal{O}(U_i K)$ cost per document $i$, where typically $U_i \ll V \ll N$.

## 5 Experiments

We first introduce our datasets and baseline models (Sections 5.1 and 5.2), then report quantitative results obtained for the three main sparse data learning problems: multi-task learning, N-shot learning and zero-shot learning (Sections 5.3 and 5.4). We also perform additional analysis on attribute-understanding tasks, robustness, and computation time (Sections 5.5-5.8).

### 5.1 Unstructured Social Activity Attribute Dataset

In previous work [11], we introduced a new attribute dataset for social activity video classification and annotation: *unstructured social activity attribute* (USAA)[9]. We selected 100 videos per-class for training and testing from 8 classes of social activities in the CCV dataset [18] (thus 1600 videos in total). We defined a wide variety of relevant attributes (illustrated in Fig. 5), and manually annotated their ground truth at the individual video level. The classes were selected as the most complex social group activities and the video length ranged from

8. This is true whether solved with variational inference [7] or MCMC [43], [27].

9. http://www.eecs.qmul.ac.uk/~yf300/USAA/download/

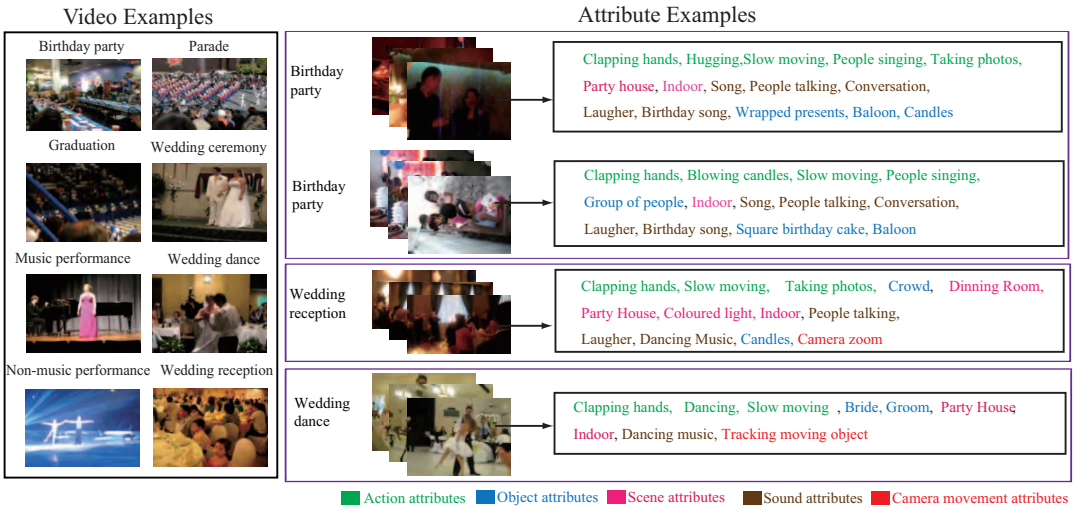


Figure 5. (Above) Different types of attributes in visual and auditory modalities are shown in different color. (Below) Examples from the eighth classes in the USAA dataset.

20 seconds to 8 minutes. The eight classes are: birthday party, graduation party, music performance, non-music performance, parade, wedding ceremony, wedding dance and wedding reception.

We experimented with two attribute-ontologies. In the first ontology, we extracted keywords from the CCV class definitions [18] and used these to obtain a set of 15 attributes. For example, the definition of graduation party is: "Graduation ceremony with *crowd*, or one or more people wearing *graduation caps* and gowns", from which we obtain attributes "crowd" and "graduation cap". In order to obtain a more exhaustive ontology of attributes, we further annotated a total of 69 attributes covering every conceivable property for this dataset including actions, objects, scenes, sounds, and camera movement. Real-world video will rarely contain such extensive tagging. However, this exhaustive annotation gives the freedom to learn on various subsets in order to quantify the effect of annotation density and biases.

Using the 69 ground-truth attributes (average density 11 per video) directly as input to a SVM, the videos can be classified with 86.9% accuracy. Individual SVM-attribute detectors achieve the mean average precision in the range $[0.22, 1]$ with average 0.785 across the entire ontology. The high variability reflects some attributes which can be detected almost perfectly (e.g., indoor scene), and others which cannot be detected given the available features (e.g., parade float). These points illustrate the challenge of these data: there is sufficient intra-class variability that even perfect knowledge of the attributes instance is insufficient for perfect classification; and moreover many attributes cannot be detected reliably.

### 5.2 Experiment Settings

For each experiment, we use 100 videos per class for testing, and a set of 100 or fewer per class for training both the attribute detectors and category classifiers. We report test set performance averaged over 5 cross-validation folds with different random selections of instances, classes, or attributes held out as appropriate. We compare the following models:

- *Direct*: Direct KNN or SVM classification on raw data without attributes. SVM is used for experiments with $> 10$ instances and KNN otherwise[10].
- *DAP*: SVM classifiers learn available UD attributes. Then zero-shot learning (ZSL) by Direct Attribute Prediction (DAP), exactly as described by [22]. It is only applicable to ZSL and deterministic attributes.
- *SVM-UD*: SVM classifiers learn available UD attributes. For N-shot learning, a logistic regression (LR) classifier then learns classes given the attribute classifier outputs[11]. This is analogous to [10]. For ZSL the SVM posteriors are matched against the manually specified prototype with NN. This is an obvious generalization of DAP [22] to non-deterministic attributes.
- *SCA*: Topic model from [39]. Learns a generative model for both class label and annotations given latent topics, in contrast to the attribute paradigm of expressing classes *in terms of* annotations/attributes. It only applies to multi-task learning.
- *ST*: Synthetic Transfer [43]. A ZSL strategy for attribute topic models: Use the source topic model to synthesize training data for novel target classes, which are then learned conventionally. We use this with our topic model. It only applies to ZSL.
- *M2LATM*: Our M2LATM is learned, then a LR classifier learns classes based on the semi-latent topic profile $\gamma$. We use 100 topics in total, with 1 UD topic

10. Our experiments show that KNN performed consistently better than SVM until #Instance >10.

11. LR is chosen over SVM because it was more robust to sparse data.

per UD attribute, 1 latent CC per class, and remaining topics are allocated to GF latent attributes.

For all experiments, we cross-validate the regularisation parameters for SVM and LR. For all SVM models, we use the $\chi^2$ kernel. For M2LATM, the user-defined part of the M2LATM topic profile $\gamma$ is estimating the same quantities as the SVM attribute classifiers, however the latter are slightly more reliable due to being discriminative classifiers, so we use these in conjunction with the latent topic profile for classification. The significance of this is quantified in Section 5.6. For semi-latent ZSL, parameter $K$ (Section 4.4) was fixed to $5\%$ of the instances.

### 5.3 Multi-task Learning

***M2LATM multi-modal latent attributes enhance multi-task learning of sparse data with incomplete ontology.*** When all classes are known in advance, shared attributes provide a mechanism for multi-task learning [33]. The statistical strength of data supporting each attribute can be aggregated across its occurrences in all classes.

Table 1 summarizes our results. We first consider the simplest upper bound scenario where the data is plentiful (100 instances per class, "100I") and the attributes are exhaustively defined (all 69UD, "A/69"). In this case all the models perform similarly except [39]. Next, we consider the sparse data and incomplete attribute space scenario of interest, with only 10 instances per class to learn from. Here Direct performs poorly due to insufficient data. Limiting the attributes to a randomly selected seven every trial ("R/7"), SVM-UD performs poorly and our M2LATM outperforms all the others by a large margin. Moreover, SVM-UD cannot apply with a completely held out attribute-ontology ("N/0"), while M2LATM performance is almost unchanged. With no attribute-ontology "N/0", SCA simplifies to supervised LDA [6][12]. Our model is thus able to share statistical strength among attributes (unlike Direct); and unlike SVM-UD, it exploits latent attributes to do so robustly to the completeness of the attribute-space definition.

***M2LATM improves both best and worst case semantic ontologies.*** In order to quantify the effectiveness of each attribute in the ontology we ranked the attributes in terms of a simple selection criteria of their "informativeness" used in text categorization [42]: Mutual information with the class (informativeness) times reliability (detection rate;). We then contrast performance between a best and worst case user-defined attribute ontology, by using the top and bottom 10% of UD attributes ("T/7" and "B/7") respectively. SVM-UD loses $14\%$ performance from the best to worst case, whereas our M2LATM model is virtually unchanged. In both cases, M2LATM provides a significant improvement over SVM-UD. SCA [39] performs significantly and consistently worse than the other models because it leverages attributes in a weaker way (as annotations rather than constraints), so we do not consider it further.

12. We used http://www.cs.princeton.edu/~chongw/slda/

| | Direct | SVM-UD | SCA[39] | M2LATM |
|---|---|---|---|---|
| 100I, A/69 | **66.0** | 65.7 | 44.0 | 65.6 |
| 10I, A/69 | 26.8 | 40.2 | 32.2 | **40.6** |
| 10I, R/7 | 26.8 | 26.4 | 25.6 | **38.3** |
| 10I, N/0 | 26.8 | - | 17.3 | **40.4** |
| 10I, T/7 | 26.8 | 32.4 | 26.0 | **38.3** |
| 10I, B/7 | 26.8 | 18.2 | 26.0 | **38.9** |

Table 1

Multi-task classification performance for USAA. 8 classes, chance = 12.5%. Row labels are I: number of training instances per class, A: all attributes, R: random subset of attributes, N: no attributes, T: top attributes, B: bottom attributes.

### 5.4 Transfer Learning

***M2LATM multi-modal latent attributes enhance N-shot learning of sparse data.*** In N-shot transfer learning, one assumes ample examples of a set of source classes, and sparse ($N$) examples of a *disjoint* set of target classes. To test this scenario, in each trial we randomly split the 8 classes into two disjoint groups of four source and target classes. We use all the data from the source task to train the attribute models (M2LATM and SVM-UD), and then use these to obtain the attribute profiles for the target task. Using the target task attribute profiles we perform N-shot learning, with the results summarized by Table 2. Importantly, SVM-UD attribute learning approach cannot deal with zero attribute situations, so can provide no benefit over Direct here, while our M2LATM improves significantly over Direct ("N/0"). In addition to drawing random subsets of attributes ("R/7" and "R/34"), we also consider the subset of 15 attributes ("O/15") we obtained from the CCV ontology (Section 5.1). Our M2LATM performs comparably or significantly better than Direct and SVM-UD in every case. Importantly M2LATM is robust to the both sparse data (performance > 35% for 1-shot learning), and exhaustiveness of the attribute-space definition (no attribute "N/0" performance within 5% of all attribute "A/69" performance). In contrast, Direct suffers strongly under sparse data 1-shot learning, and SVM-UD suffers with sparse attribute-space (7UD "R/7" performance 12% below all attribute performance). The robust performance of M2LATM is enabled by the semi-latent attribute representation.

***M2LATM multi-modal latent attributes enhances zero-shot learning.*** Like N-shot learning, the task is to learn transferrable attributes from a source dataset for use on a disjoint target set. Instead of providing training examples, users manually specify the definition of each novel class in the user-defined attribute space. ZSL is often evaluated in simple situations where classes have unique 1:1 definitions in the attribute-space [22]. For unstructured social data [18], strong intra-class variability violates this assumption, making evaluation more subtle. To define the novel classes, we take the thresholded mean (as in [10], [11]) of the attribute profiles for each instance of that class from our ground-truth.

Our results are summarized in Table 3. The key ob-

| | 1-shot | | | 5-shot | | | 10-shot | | |
|---|---|---|---|---|---|---|---|---|---|
| | Direct | SVM-UD | M2LATM | Direct | SVM-UD | M2LATM | Direct | SVM-UD | M2LATM |
| N/0 | 29.0 | - | **35.3** | 33.6 | - | **48.0** | 35.7 | - | **53.0** |
| R/7 | 29.0 | 30.9 | **35.9** | 33.6 | 36.9 | **48.0** | 35.7 | 38.7 | **52.2** |
| R/34 | 29.0 | 35.0 | **36.5** | 33.6 | 44.5 | **48.9** | 35.7 | 47.5 | **52.8** |
| O/15 | 29.0 | 36.1 | **37.7** | 33.6 | 46.8 | **49.7** | 35.7 | 50.2 | **53.3** |
| A/69 | 29.0 | **39.1** | 38.6 | 33.6 | 49.7 | **52.1** | 35.7 | 52.5 | **56.1** |

Table 2
N-shot classification performance for USAA dataset (4v4 classes, chance = 25%) .

| | SVM-UD | ST[43] | M2LATM |
|---|---|---|---|
| R/7 | 27.1 | 18.1 | **33.8** |
| O/15 | 31.3 | 36.9 | **39.4** |
| R/34 | 36.7 | 30.9 | **39.2** |
| A/69 | 33.2 | 31.0 | **41.9** |

Table 3
Zero-shot classification performance (%) for USAA (4v4 classes, chance = 25%).

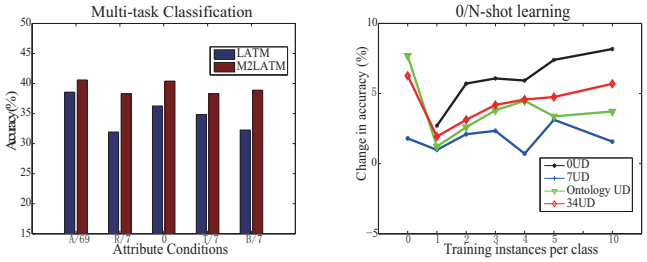


Figure 6. Exploiting multi-modality: LATM vs M2LATM for USAA dataset. Left: Multi-task classification. Right: 0/N-shot learning shown as margin of M2LATM over LATM – positive value means increase of accuracy.

| Modality | Attributes |
|---|---|
| Static (SIFT) | Candles, Dark outdoors, Party House |
| Motion (STIP) | Slow moving, Crowd, Bright outdoors |
| Audio (MFCC) | Laughter, Singing, Instrumental music |
| Static+Motion | Hold microphone, Birthday caps, Crowd |
| Static+Audio | Singing, People in a row, Fast moving |
| Audio+Static | Formal speech, Crowds, Dining room |

Table 4
Top-3 attributes most strongly associated with modalities.

servation is that using latent attributes to support the user-defined attributes (Section 4) allows M2LATM to improve on SVM-UD [22], which only uses UD attributes in ZSL. This is a surprising and significant result, because it is not obvious that ZSL from human description should be able to be exploit latent data-driven attributes. Additionally, we compare the synthetic data transfer strategy from [43], generating $N = 50$ synthetic data instances per class from the zero-shot definition, and training the classifier based on the learned profiles for these. We found that this underperformed DAP in most cases, and M2LATM in every case. This is unsurprising, because synthetic data adds no truly new information: it is generated from the UD word-topic distributions $\boldsymbol{\beta}$, learned from the source dataset. M2LATM already uses $\boldsymbol{\beta}$, but additionally exploits latent topics.

### 5.5 Attribute Understanding

***M2LATM makes effective use of multiple modalities.*** An important contribution of M2LATM is explicitly representing the correspondence between attributes and features of each modality, bridging the cross-modal gap. Existing approaches often ignore this issue either by using only one modality [43] or taking a weighted average/concatenation [22] of modalities, which introduces issues in selection of scaling/weighting factors. We compare M2LATM against a simpler variant of our approach approach, LATM. LATM takes the standard approach of simply concatenating feature vectors (with rescaling to ensure modalities are represented on the same scale). Explicit multi-modality consistently improves the results relative to simple concatenation in multi-task (Fig. 6, left) and transfer (Fig. 6, right) learning.

***M2LATM can associate attributes with their observation modality.*** To provide further insight into the capabilities of our cross-modal model, we consider a novel task of learning which modalities each attribute appears in. This can be computed from the relative proportion of words assigned by the model to static appearance, motion or auditory modalities when explaining a given topic/attribute. That is, comparing modalities $m$ in $\sum_i \gamma_{ikm}$ for each attribute $k$. To illustrate this, Table 4 reports the top-3 attributes most strongly associated with each modality and each modality pair (as assessed by geometric mean). Clearly most attributes have associations with intuitive modalities.

***M2LATM can detect semantically surprising multimedia content.*** As a final example of attribute understanding, we illustrate some examples of surprising semantics discovered by our framework – based on the correlations encoded in the class-attribute relationships (Section 4.3.5). Fig. 7(A) is correctly classified as a birthday party. However, both the "*instrumental music*" (auditory) and "*musical instruments*" (static appearance) attributes are detected (a person sings "happy birthday" using a guitar), which are unusual in birthday party settings. Fig. 7(B) is a music performance video, which unexpectedly has the "*costume*" attribute, as there are also costumed actors on stage. A wedding ceremony is shown in Fig. 7(C), where guests are unusually drinking


Figure 7. Examples of surprising videos: (A) birthday party with instrumental music, (B) music performance with costumes, (C) wedding ceremony with drinking glasses, (D) an indoor parade.

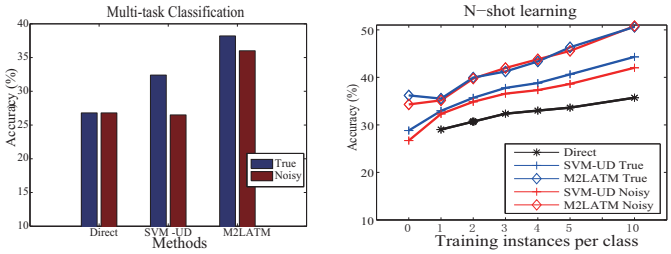


Figure 8. Robustness to attribute label-noise in multi-task classification and zero/N-shot learning.

during the ceremony ("*drinking glass*" attribute). Fig. 7(D) illustrates an example of expected attributes which are surprisingly absent. In this case the video is correctly classified as a parade, however the expected attributes "bright outdoor scene" and "parade float" are absent because it is, unusually, an indoor parade.

### 5.6 Further Evaluations

***M2LATM improves robustness to label noise.*** An important challenge for learning from real-world user data, or AMT annotations, is dealing with label-noise. We expect our model to deal better with label noise in the user-defined attributes, because it can additionally leverage automatically discovered latent attributes for a more robust overall representation. To simulate this, we repeated the previous multi-task and zero/N-shot learning experiments, but randomly flipped $50\%$ of attribute bits on $50\%$ of the training videos (so $25\%$ wrong annotations). M2LATM is more robust than SVM-UD (Fig. 8 red vs blue), sometimes dramatically so. For example, when subject to label noise, multi-task classification performance of SVM-UD drops 8% (vs only 3% for M2LATM) and actually performs worse than Direct. ***User-defined and latent attributes should be learned jointly.*** M2LATM model has three complementary types of topics that define the semi-latent attribute space. An

| | 1-shot | | 5-shot | | 10-shot | |
|---|---|---|---|---|---|---|
| | Ind. | Joint | Ind. | Joint | Ind. | Joint |
| R/7 | 35.0 | **38.8** | 42.0 | **48.0** | 48.0 | **52.2** |
| A/69 | **38.8** | 38.6 | 49.4 | **52.1** | 54.6 | **56.1** |

Table 5
Independent vs joint learning of semi-latent attributes. N-shot transfer. (4v4 classes, chance = 25%) .

advantage of our model is to learn these jointly. To quantify this, we also learn them separately by training a batch of SVM classifiers (for UD topics), a constrained topic model (just CC topics), and an unsupervised topic model (GF topics). We compare performance using the concatenated output of the individual models vs the output of the jointly model in N-shot transfer learning. The results (Table 5) show that joint learning is always similar or significantly better than independent learning, so joint learning of latent attributes is indeed important to ensure they learn complementary aspects to UD attributes.

***Significance of using SVM posteriors as user-defined attributes.*** We use M2LATM to jointly learn UD, CC and GF attributes in a single generative model, with the aim of ensuring that latent attributes are complementary to user-defined attributes. However, as discussed in Section 5.2, we ultimately use the SVM posteriors in place of the UD topics because, being discriminatively trained strong classifiers, they perform slightly better. However, this is not a significant factor in our model's performance: across all the experiments, the margin of using SVM attribute classifiers over topic posteriors is $[-3\% \sim 4\%]$.

### 5.7 Analysis of Discovered Latent Attributes

***Latent attributes can discover user-defined attributes from a withheld ontology, as well as novel attributes outside the full ontology.*** In this section we investigate what is learned by latent attributes: can they discover UD attributes not provided in the ontology, and do they discover anything outside of the full UD ontology? Firstly, we define the distance between learned attributes $i$ and $j$ as the normalized correlation between their multinominal parameters $D(i,j) = \beta_i^T \beta_j / (\|\beta_i\| \, \|\beta_j\|)$. Fig. 9 shows the sorted similarity matrix between attributes for M2LATM learned in a conventional A/69 and semi-latent R/7 attribute setting. The diagonal structure shows that latent attributes have largely discovered many of the semantic UD attributes of interest to users. The uncorrelated strip to the right represents latent attributes in the R/7 model which have discovered aspects of the data not covered by the UD attributes.

To visualize an attribute, we select its top-N most likely words (from $\beta$), and then plot occurrences of these words on videos with high probability for this attribute ($\gamma$). Fig. 10 (top row) shows an example of static appearance (SIFT) attributes *bride* and *cake*. The high degree of overlapping between red circles and red crosses indicates that the re-discovered latent attributes

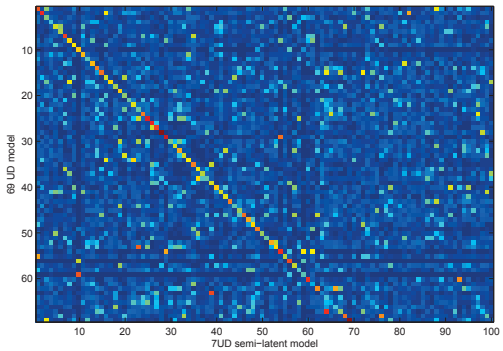


Figure 9. Similarity between user-defined and latent attributes.

match the withheld UD attributes well. Examples of STIP attributes *blow candle*, and *dancing* are shown in Fig. 10 (second row). For auditory attributes, we show *birthday song* and *speech* in Fig. 10 (third row). In this case, we plot the time-series of the attribute weight for the corresponding UD attribute and the latent attribute which rediscovered it along with ground-truth for when the particular sound was audible. All of these latent-attributes were GF type, except *birthday song*, which was CC – being uniquely selective for birthday-party class.

Finally, to further illustrate the value of latent attributes, we visualized some latent attributes with no similarity to any existing UD attribute (i.e., those on the right strip of Fig. 9). This revealed new attributes which we had not included in our ontology despite intending it to be exhaustive. Fig. 10 (bottom row) shows two examples: (i) a *horizontal line* attribute, which the model learns is informative for classes with stages and fences such as concerts and performances; and (ii) a *tree* attribute, which the model learns is informative for typically outdoor or situations such as wedding receptions and parades. These results support our motivating point that manual ontologies are almost certainly *incomplete*, and benefit from being complemented with a set of latent attributes.

### 5.8 Computational Scalability

In Section 4.5, we introduced a new sparse vocabulary-domain representation of our inference algorithm. To contrast the improved scalability of this representation (Eq. (9)) vs. the standard word-domain approach (Eq. (4), also used by [43], [39], [29], [11]), we recorded the matlab computation time for 10 instance multi-task learning on the USAA data. Our model required 30 minutes versus to 5 hours for the conventional approach. This margin grows with the video length and density of features, so this is an important contribution for scalability.

### 5.9 Experiments on Animals with Attributes (AwA)

Our model is not specific to videos/social activities. We also study the well known AwA dataset, (see [22] for full details). AwA dataset defines 50 classes of animals, and 85 associated attributes (such as *furry*, and *has claws*).

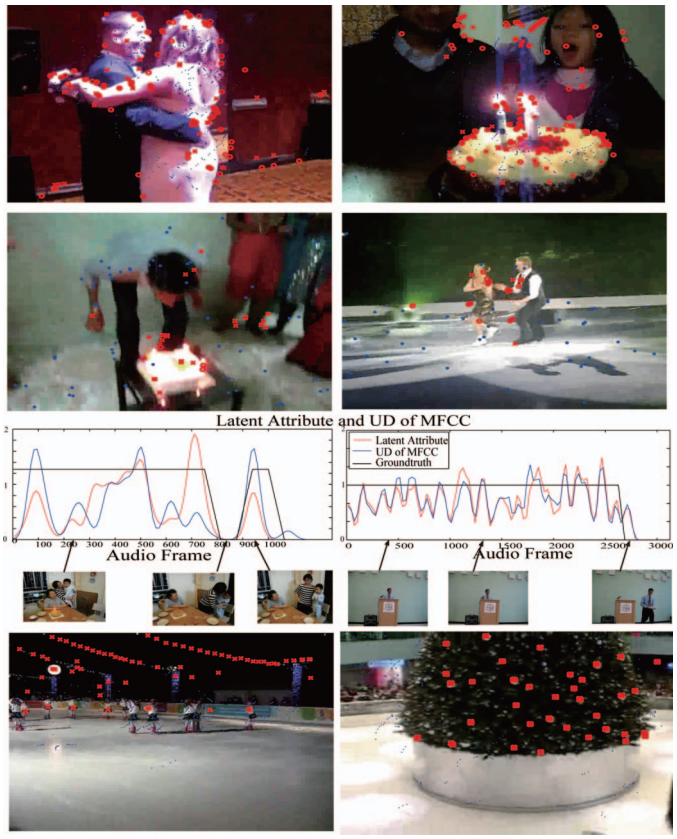


Figure 10. Visualization of user-defined (circles) and corresponding latent attributes (crosses). Red circles illustrate representative words from the UD attribute (A/69); red crosses illustrate the words from the corresponding latent attribute which discovered these concepts when withheld (R/7). Blue dots illustrate interest-points not related to attributes concerned.

There are 30475 images with at least 92 examples of each class. We use the same six BoW features from [22]. In contrast to USAA dataset, each class has a distinct deterministic definition in terms of attributes. For M2LATM, we keep the complexity fixed at 150 topics: with 1CC attribute per class, up to 85 user-defined attributes, and the others are GF latent attributes. There are six different kind of features extracted to describe the AwA images.

Table 6 shows N-shot learning results for AwA, with the attributes learned from all instances of 40 classes, and the target task learned from $1-10$ instances of the held out ten classes (same condition as [22]). The same general results hold: M2LATM performs comparably or better than the others in most cases. Notably, although SVM-UD slightly outperforms M2LATM with the exhaustive A/85 condition (due to M2LATM's larger number of dimensions over fitting slightly), the use of latent attributes enables M2LATM to outperform SVM-UD in the most relevant and challenging cases of few UD attributes.

The ZSL results are shown in Table 7. Here, because the AwA attributes are deterministic, we were able to implement and apply DAP zero-shot learning precisely as described in [22]. Different from [22], we found that attribute priors provided a noticeable improvement in

| | 1-shot | | | 5-shot | | | 10-shot | | |
|---|---|---|---|---|---|---|---|---|---|
| Condition | Direct | SVM-UD | M2LATM | Direct | SVM-UD | M2LATM | Direct | SVM-UD | M2LATM |
| N/0 | 16.4 | - | **19.2** | 21.5 | - | **30.5** | 23.6 | - | **35.9** |
| R/9 | 16.4 | 25.1 | **27.1** | 21.5 | 32.6 | **35.6** | 23.6 | 36.4 | **39.0** |
| R/42 | 16.4 | **30.7** | 28.3 | 21.5 | 42.5 | **42.5** | 23.6 | 45.0 | **45.7** |
| A/85 | 16.4 | **31.9** | 28.5 | 21.5 | **43.4** | 38.0 | 23.6 | **46.8** | 43.0 |

Table 6
N-shot classification performance for AwA dataset (40v10 classes, chance = 10%).

| | | No Attrib Prior | | Attrib Prior | |
|---|---|---|---|---|---|
| | [43]/[26]/[20] | DAP | M2LATM | DAP | M2LATM |
| R/9 | - | 26.3 | 26.9 | 27.8 | **29.2** |
| R/42 | - | 34.4 | 38.2 | 36.0 | **39.7** |
| A/85 | 33.0/33.0/32.7 | 37.0 | 39.2 | 39.2 | **41.3** |

Table 7
Zero-shot classification performance (%) for AwA (40v10 classes, chance = 10%).

performance, so we show results with and without priors. In general, M2LATM outperforms DAP across the range of ontology completeness. For context, we also show the $\approx 33\%$ figure reported by several recent ZSL studies [43], [26] and [20], although these conditions may not be exactly comparable to ours. This highlights the fact that our approach outperforms very recent methods with as few as half of the available attributes (R/42).

## 6 Conclusion

### 6.1 Summary

In this paper we developed a new framework for multimedia understanding focused on bridging the semantic and cross-modal gaps via an attribute-learning approach. We address the limitations of previous studies including reliance on an exhaustive manual specification of the attribute-space, ignoring or simplistically dealing with multi-modal content, and the unrealistic requirement of noiseless annotation of attributes. In particular, we are able to: (i) flexibly learn a full semantic-attribute space whether exhaustively defined, completely unavailable, available in a small subspace (i.e., present but sparse), or available but with noisy examples; (ii) improve multi-task and N-shot learning by leveraging latent attributes; (iii) go beyond existing zero-shot learning approaches (which only use user-defined attributes) by also exploiting latent attributes; (iv) explicitly leverage attributes in conjunction with multi-modal data to improve cross-media understanding, enabling new tasks such as explicitly learning which modalities particular attributes appear in; and (v) make our topic model applicable to large multimedia data by expressing it in a significantly more scalable way than previous studies – invariant to the length of the input video and density of the features.

### 6.2 Future Work

There remain a number of important open questions to be addressed. Thus far, our attribute-learner does not consider inter-attribute correlation explicitly [32], [36], though this limitation is shared by most other attribute learners with the exception of [24]. For our task, this can be addressed straightforwardly by generalizing the correlated topic model (CTM) [5] instead of the conventional LDA [7], which should produce commensurate gains in performance to those observed elsewhere [24].

The complexity of our model in terms of the size of the attribute/topic-space was fixed to a reasonable value throughout, and we focused on learning with attribute-constraints on the topics. A more desirable solution would be a non-parametric framework which could infer the appropriate dimension of the latent attribute-space automatically given available UD attributes.

## References

[1] A. Asuncion, M. Welling, P. Smyth, and Y. W. Teh. On smoothing and inference for topic models. In *Uncertainty in Artificial Intelligence*, 2009.
[2] H. Barrow and J. Tenenbaum. Recovering intrinsic scene characteristics from images. *Computer Vision Systems*, 1978.
[3] I. Biederman. Recognition by components - a theory of human image understanding. *Psychological Review*, 1987.
[4] D. M. Blei and M. I. Jordan. Modeling annotated data. In *Proceedings of the International ACM SIGIR conference on Research and Development in Information Retrieval*, pages 127–134, 2003.
[5] D. M. Blei and J. Lafferty. A correlated topic model of science. *Annals of Applied Statistics*, 1:17–35, 2007.
[6] D. M. Blei and J. McAuliffe. Supervised topic models. In *Neural Information Processing Systems*, 2007.
[7] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.
[8] L. Cao, S.-F. Chang, N. Codella, C. Cotton, D. Ellis, L. Gong, M. Hill, G. Hua, J. Kender, M. Merler, Y. Mu, A. Natsev, and J. R. Smith. Ibm research and columbia university trecvid-2011 multimedia event detection (med) system. In *NIST TRECVID Workshop*, 2011.
[9] A. Farhadi, I. Endres, and D. Hoiem. Attribute-centric recognition for cross-category generalization. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2352–2359, 2010.
[10] A. Farhadi, I. Endres, D. Hoiem, and D. Forsyth. Describing objects by their attributes. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2009.
[11] Y. Fu, T. Hospedales, T. Xiang, and S. Gong. Attribute learning for understanding unstructured social activity. In *European Conference on Computer Vision*, 2012.
[12] A. Hauptmann, R. Yan, W.-H. Lin, M. Christel, and H. Wactlar. Can high-level concepts fill the semantic gap in video retrieval? a case study with broadcast news. *Multimedia, IEEE Transactions on*, 9(5):958 –966, aug. 2007.
[13] T. Hospedales, S. Gong, and T. Xiang. Learning tags from unsegmented videos of multiple human actions. In *International Conference on Data Mining*, 2011.
[14] T. Hospedales, S. Gong, and T. Xiang. Video behaviour mining using a dynamic topic model. *International Journal of Computer Vision*, 2011.

[15] T. Hospedales, J. Li, S. Gong, and T. Xiang. Identifying rare and subtle behaviours: A weakly supervised joint topic model. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2011.

[16] S. J. Hwang, F. Sha, and K. Grauman. Sharing features between objects and their attributes. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2011.

[17] Y.-G. Jiang, J. Yang, and C.-W. Ngo. Representations of keypoint-based semantic concept detection: a comprehensive study. *IEEE Transaction on Multimedia*, 2008.

[18] Y.-G. Jiang, G. Ye, S.-F. Chang, D. Ellis, and A. C. Loui. Consumer video understanding: A benchmark database and an evaluation of human and machine performance. In *ACM International Conference on Multimedia Retrieval*, 2011.

[19] Y.-G. Jiang, X. Zeng, G. Ye, S. Bhattacharya, D. Ellis, M. Shah, and S.-F. Chang. Columbia-ucf trecvid2010 multimedia event detection: Combining multiple modalities, contextual concepts, and temporal matching. In *NIST TRECVID Workshop*, 2010.

[20] P. Kankuekul, A. Kawewong, S. Tangruamsub, and O. Hasegawa. Online incremental attribute-based zero-shot learning. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2012.

[21] N. Kumar, A. C. Berg, P. N. Belhumeur, and S. K. Nayar. Attribute and simile classifiers for face verification. In *IEEE International Conference on Computer Vision*, 2009.

[22] C. H. Lampert, H. Nickisch, and S. Harmeling. Learning to detect unseen object classes by between-class attribute transfer. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2009.

[23] I. Laptev. On space-time interest points. *International Journal of Computer Vision*, 64:107–123, September 2005.

[24] J. Liu, B. Kuipers, and S. Savarese. Recognizing human actions by attributes. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2011.

[25] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vision*, 60, 2004.

[26] D. Mahajan, S. Sellamanickam, and V. Nair. A joint learning framework for attribute models and object descriptions. In *IEEE International Conference on Computer Vision*, pages 1227–1234, 2011.

[27] D. Mimno, H. M. Wallach, J. Naradowsky, D. A. Smith, and A. McCallum. Polylingual topic models. In *Conference on Empirical Methods on Natural Language Processing*, 2009.

[28] D. Newman, A. Asuncion, and P. Smyth. Distributed algorithms for topic models. *Journal of Machine Learning Research*, 10:1801–1828, 2009.

[29] J. C. Niebles, H. Wang, and L. Fei-Fei. Unsupervised learning of human action categories using spatial-temporal words. *International Journal of Computer Vision*, 79(3):299–318, 2008.

[30] M. Palatucci, G. Hinton, D. Pomerleau, and T. M. Mitchell. Zero-shot learning with semantic output codes. In *Neural Information Processing Systems*, 2009.

[31] D. Parikh and K. Grauman. Interactively building a discriminative vocabulary of nameable attributes. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2011.

[32] G.-J. Qi, X.-S. Hua, Y. Rui, J. Tang, T. Mei, and H.-J. Zhang. Correlative multi-label video annotation. In *ACM International Conference on Multimedia*, 2007.

[33] R. Salakhutdinov, A. Torralba, and J. Tenenbaum. Learning to share visual appearance for multiclass object detection. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2011.

[34] C. G. M. Snoek, B. Huurnink, L. Hollink, M. de Rijke, G. Schreiber, and M. Worring. Adding semantics to detectors for video retrieval. *IEEE Transactions on Multimedia*, 9:975–986, 2007.

[35] A. Sorokin and D. Forsyth. Utility data annotation with amazon mechanical turk. In *CVPR Workshops*, 2008.

[36] J. Tang, X.-S. Hua, M. Wang, Z. Gu, G.-J. Qi, and X. Wu. Correlative linear neighborhood propagation for video annotation. *IEEE Transactions on Systems, Man, and Cybernetics, Part B*, 39(2):409–416, 2009.

[37] J. Tang, S. Yan, R. Hong, G.-J. Qi, and T.-S. Chua. Inferring semantic concepts from community-contributed images and noisy tags. In *ACM International Conference on Multimedia*, 2009.

[38] G. Toderici, H. Aradhye, M. Pasca, L. Sbaiz, and J. Yagnik. Finding meaning on youtube: Tag recommendation and category discovery. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3447–3454, 2010.

[39] C. Wang, D. Blei, and F.-F. Li. Simultaneous image classification and annotation. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2009.

[40] Y. Wang and G. Mori. Human action recognition by semilatent topic models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(10):1762–1774, 2009.

[41] K. Yang, X.-S. Hua, M. Wang, and H.-J. Zhang. Tag tagging: Towards more descriptive keywords of image content. *Multimedia, IEEE Transactions on*, 13:662 –673, 2011.

[42] Y. Yang and J. O. Pedersen. A comparative study on feature selection in text categorization. In *International Conference on Machine Learning*, 1997.

[43] X. Yu and Y. Aloimonos. Attribute-based transfer learning for object categorization with zero/one training example. In *European Conference on Computer Vision*, 2010.

[44] X. Zhu. Semi-supervised learning literature survey. Technical Report 1530, University of Wisconsin-Madison Department of Computer Science, 2007.

**Yanwei Fu** received the BSc degree in information and computing sciences from Nanjing University of Technology in 2008; and the MEng degree in the Department of Computer Science & Technology at Nanjing University in 2011, China. He is now pursuing his PhD in vision group of EECS, Queen Mary University of London. His research interest is attribute learning, topic model, learning to rank, video summarization and image segmentation.

**Timothy Hospedales** received the Ph.D degree in Neuroinformatics from University of Edinburgh in 2008 and now is a Lecturer (Assistant Professor) in Computer Science at Queen Mary University of London. His research interests include probabilistic modelling and machine learning applied variously to problems in computer vision, data mining, interactive learning and neuroscience. He has published over 20 papers in major international journals and conferences.

**Tao Xiang** received the PhD degree in electrical and computer engineering from the National University of Singapore in 2002. He is a currently a Senior Lecturer (Associate Professor) in the School of Electronic Engineering and Computer Science, Queen Mary University of London. His research interests include computer vision, statistical learning, video processing, and machine learning, with a focus on interpreting and understanding human behaviour. He has published over 100 papers on international journals and conferences and coauthored a book Visual Analysis of Behaviour: From Pixels to Semantics.

**Shaogang Gong** is Professor of Visual Computation at Queen Mary University of London, a Fellow of the Institution of Electrical Engineers and a Fellow of the British Computer Society. He received his D.Phil in computer vision from Keble College, Oxford University in 1989. His research interests include computer vision, machine learning and video analysis.