Attribute Learning for Understanding Unstructured Social Activity

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

School of EECS, Queen Mary University of London, UK
{yanwei.fu,tmh,txiang,sgg}@eecs.qmul.ac.uk

Abstract. The rapid development of social video sharing platforms has created a huge demand for automatic video classification and annotation techniques, in particular for videos containing social activities of a group of people (e.g. YouTube video of a wedding reception). Recently, attribute learning has emerged as a promising paradigm for transferring learning to sparsely labelled classes in object or single-object short action classification. In contrast to existing work, this paper for the first time, tackles the problem of attribute learning for understanding group social activities with sparse labels. This problem is more challenging because of the complex multi-object nature of social activities, and the unstructured nature of the activity context. To solve this problem, we (1) contribute an unstructured social activity attribute (USAA) dataset with both visual and audio attributes, (2) introduce the concept of semi-latent attribute space and (3) propose a novel model for learning the latent attributes which alleviate the dependence of existing models on exact and exhaustive manual specification of the attribute-space. We show that our framework is able to exploit latent attributes to outperform contemporary approaches for addressing a variety of realistic multi-media sparse data learning tasks including: multi-task learning, N-shot transfer learning, learning with label noise and importantly zero-shot learning.

1 Introduction

With the rapid development of digital and mobile phone cameras and proliferation of social media sharing, billions of unedited and unstructured videos produced by consumers are uploaded to the social media websites (e.g. YouTube) but few of them are labelled. Obtaining exhaustive annotation is impractically expensive. This huge volume of data thus demands effective methods for automatic video classification and annotation, ideally with minimised supervision. A solution to these problems would have huge application potential, e.g., content-based recognition and indexing, and hence content-based search, retrieval, filtering and recommendation of multi-media..

In the paper, we tackle the problem of *automatic classification and annotation of unstructured group social activity*. Specifically, we are interested in home videos of social occassions such graduation ceremony, birthday party, and wedding reception which feature activities of group of people ranging anything

A. Fitzgibbon et al. (Eds.): ECCV 2012, Part IV, LNCS 7575, pp. 530–543, 2012.

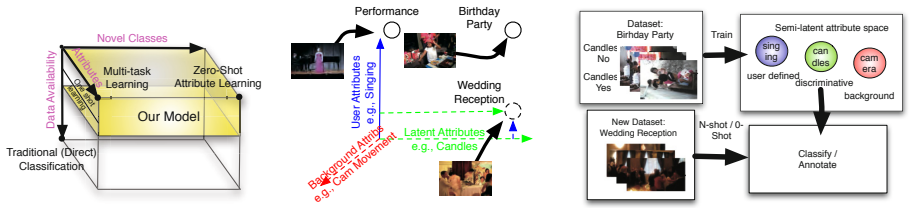
Fig. 1. Examples in social activity attribute video dataset. Different types of attributes of both visual and audio modalities are shown in different colour.

between a handful to hundreds (Fig. 1). By classification, we aim to categorise each video into a class; and by annotation we aim to predict what are present in the video. This implies a wide range of multi-modal annotation types including object (e.g. group of people, cake, balloon), action (e.g. clapping hands, hugging, taking photos), scene (e.g. indoor, garden, street), and sound (e.g. birthday song, dancing music). We consider that the problem of classification and annotation are inter-related and should be tackled together. There have been extensive works on image classification and annotation [1]. However, little effort has been taken on video data, especially on unstructured group social activity video.

We propose to solve the problem using an attribute learning framework, where annotation becomes the problem of attribute prediction and video classification is helped by a learned attribute model. Attributes describe the characterisitics that embody an instance or a class. Recently, attribute-based learning [2,3,4,5,6] has emerged as a powerful approach for image and video understanding. Essentially attributes answer the question of *describing* a class or instance in contrast to the typical (classification) question of *naming* an instance [2,3]. The attribute description of an instance or category is useful as a semantically meaningful intermediate representation to bridge the gap between low level features and high level classes [6]. Attributes thus facilitate transfer and zero-shot learning [6] to alleviate issues of the lack of labelled training data, by expressing classes in terms of well known attributes.

We contribute a new benchmarking multi-modal attribute dataset for social activity video classification and annotation: *unstructured social activity attribute* (USAA) dataset[1]. It comprises of 8 classes (around 1500 videos totally) and the visual and audio content of each video is manually annotated using 69 multi-modal binary attributes. Figure 1 shows examples of videos with annotated

[1] Downloadable from `http://www.eecs.qmul.ac.uk/~yf300/USAA/download/`



(a) Problem Context (b) Semi-latent Attribute Space (c) Model Overview

Fig. 2. (a) Our approach to semi-latent attribute space learning can be applied in various problem contexts. (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 and latent attributes. (c) Model overview. New classes are learned via expression in terms of learned semi-latent attribute-space from (b).

attributes. Learning these attributes can support a wide range of studies including object recognition, scene classification, action recognition and audio event recognition. There are a number of unique characters and challenges of this dataset which can be beneficial to the wide community: (1) The data is weakly labelled (each attribute annotation does not tell which part of the video contribute to that attribute). (2) Different instances of one social activity video class (Fig. 1) typically cover a wide variety of attributes (e.g., birthday party class may or may not exhibit candles). One thus cannot make the assumption that a class can be uniquely determined by a deterministic vector of binary attributes [2]. (3) Even with 69 attributes, one cannot assume that the user-defined space of attributes is perfectly and exhaustively defined due to limited annotation, and subjectiveness of manual annotation. (4) The most semantically salient attributes may not be the most discriminative and most discriminative attributes may not correspond to semantic concept and thus can never be manually defined. Discovering and learning those discriminative yet latent attributes thus becomes the key.

To this end, in this paper we introduce the novel concept of semi-latent attribute space. As illustrated in Fig. 1(b), this attribute space consists of three types of attributes: user-defined (UD) attributes, class-conditional (CC) discriminative latent attributes and background non-discriminative (BN) latent attributes. Among the two types of latent attributes, the CC attributes are discriminative attributes which are predictive of class, whilst the BN attributes are uncorrelated to class of interest and should thus be ignored as background data, e.g. random camera or background object movements which are common characteristics of most unstructured social activity videos. It is crucial that these three types of attributes should be learned jointly so that the CC attributes do not repeat the user-defined attributes (UD attributes often are also discriminative), and are separated explicitly from background attributes which explain away irrelevant dimension of the data [7].

To learn this semi-latent attribute space, we present a new approach to attribute learning based on a probabilistic topic model [8,9]. A topic model is chosen because it provides an intuitive mechanism for modelling latent attributes using latent topics. We consider the attribute/topic learning process as *semantic feature reduction* [6] from the raw data to a lower dimensional attribute space (where the axes are the attribute/topic set) (Fig. 1(b)). Classification is then performed in this semantic feature space. To learn the three types of attributes: UD, CC, and BN, the topic model learns three types of topics, namely UD topics, CC topics and BN topics. Among them the UD topics are learned supervised using the labelled use-defined attributes, whilst the learning of CC is supervised by the class label available during training, and the BN topics are learned unsupervised. An important advantage of this approach is that it can seamlessly bridge the gap between context where the attribute space is completely and precisely specified by the user; and scenarios where the attribute space is completely unknown (Fig. 1(a)). This means that unlike existing approaches, our approach is robust to the amount of domain knowledge / annotation budget possessed by the user. Specifically, if the relevant attribute space is exhaustively and correctly specified, we create a topic or set of topics for each attribute, and learn a topic model where the topics for each instance are constrained to not violate the instance-attribute labels. However, if the attribute space is only partially known, we complete the semantic space using *latent* attributes by learning two additional types of topics: CC topics to discover unique attributes of each known class [9]; and BN topics to segment out background non-discriminative attributes [7]. At the extreme, if the relevant attribute space is completely unknown, the latent attributes alone can discover a discriminative and transferrable intermediate representation. Figure 1(c) gives an overview of the process.

2 Related Work

Learning attribute-based semantic representations of data has recently been topical for images [2,5,10,4,11]. The primary contribution of attribute-based representations has been to enable transfer learning (via attribute classifiers) to learn classes with few or zero instances. However, most of these studies [2,5,4,11] assume that an exhaustive space of attributes has been manually specified. Moreover, it is also assumed that each class is simple enough to be determined by a single list of attributes. In practice a complete space of relevant attributes is unlikely to be available a priori since human labelling is limited and the space of classes is unbounded. Furthermore, semantically obvious attributes for humans do not necessarily correspond to the space of useful and computable discriminative attributes [12] (Fig. 1(b)).

A few studies ([3] for object and [13] for action) have considered augmenting user-defined (UD) attributes with data-driven attributes which correspond to our definition of class-conditional (CC) attributes. However these do not span the full spectrum between unspecified and fully specified attribute-spaces as cleanly as our model. Notably, they learn UD attributes and CC attributes separately.

This means that the learned CC attributes are not necessarily complementary to user-defined ones (i.e., they may be redundant). Additionally, some data-driven attributes may be irrelevant to other discriminative tasks, and should thus be ignored. This may not be a problem for annotating an object bounding box [3] and a single object action without people interaction [13] where background information does not present a big issue for learning discriminative foreground attributes. It is however a problem for unstructured social activity video where shared characteristics (therefore attributes) across classes may not be relevant for either classification or annotation. In our approach, by jointly learning user-defined, class-conditional and background non-discriminative (BN) attributes, we ensure that the latent attribute space is both complementary and discriminative.

Probabilistic topic models [8] have been used quie extensively in modelling images [1] and video [14,15,9,7]. However, the topic spaces in those models are used for completely unsupervised dimensionality reduction. Here, we focus on an attribute learning interpretation to learn a semantically meaningful semi-latent topic-space, which leverages as much from any given prior knowledge, either in the form of sparely labelled either class or user-defined attributes.

User-defined video attribute learning is related to the video concept detection (video ontology) work in the multimedia community [16,17,18,19,20,21,22,23] which has defined top-down shared visual concepts, in order to recognise them in video. There are several TRECVID challenges about video ontologies, e.g. in TRECVID Multimedia Event Detection [2]. However, these studies generally consider strongly labelled data and prescriptive ontologies and do not leverage discriminative latent attributes for classification.

This paper makes the following specific contributions: (i) To study the issue of unstructured group social activity video classification and annotation, we present a multi-modal social activity attribute dataset to be made available to the community. (ii) We propose a new topic-model based approach for attribute learning. By learning a unified semi-latent space of user-defined and two types of latent-attributes, we are able to learn a complete and discriminative attribute-space in a way that is robust to any amount of user prior-knowledge. (iii) We show how these properties improve a variety of tasks in the sparse data domain including multi-task learning, N-shot and 0-shot transfer learning. (iv) Our unified framework enables us to leverage latent attributes even in zero-shot learning which has not been attempted before.

3 Methods

3.1 Formalisation

Context. Prior work on detection or classification typically takes the approach of learning a classifier $F : \mathcal{X}^d \rightarrow \mathcal{Z}$ mapping d-dimensional raw data $\mathcal{X}$ to label

[2] `http://www.nist.gov/itl/iad/mig/med12.cfm`

$\mathcal{Z}$ from training data $D = \{(\mathbf{x}_i, z_i)\}_{i=1}^n$. A variant of the standard approach considers a composition of two mappings:

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

where L maps the raw data to an intermediate representation $\mathcal{Y}^a$ (typically with $a < d$) and then S maps the intermediate representation to the final class $\mathcal{Z}$. Examples of this approach include dimensionality-reduction via PCA [24] (where L is learned to explain the variance of $\mathbf{x}$) or linear discriminants and multi-layer neural networks (where L is learned to predict $\mathcal{Z}$).

Attribute learning [2,6] exploits the idea of manually defining $\mathcal{Y}$ as a *semantic feature* or *attribute* space. L is then learned by direct supervision with pairs of instances and attribute vectors $D = \{(\mathbf{x}_i, \mathbf{y}_i)_{i=1}^n\}$. A key feature of this approach is that it permits practical zero-shot learning: the recognition of novel classes without training examples $F : X^d \rightarrow \mathcal{Z}^*$ ($\mathcal{Z}^* \notin \mathcal{Z}$) via the learned attribute mapping L and a manually specified template $S^*$of the novel class. Attribute learning can also assist general multi-task and N-shot transfer learning, where we learn a second "target" dataset $D^* = \{(\mathbf{x}_i, z_i^*)\}_{i=1}^m$but $m \ll n$. Here, the attribute mapping L is learned from the large "source" dataset, and is transferred to the target task, leaving only parameters of S to be learned. Most prior attribute-learning work, however, assumes the semantic space $\mathcal{Y}^a$ is completely defined in advance, an assumption we would like to relax.

Semi-latent Attributes. We aim to define an attribute-learning model L which can learn an attribute-space $\mathcal{Y}^a$ from training data D where $|\mathbf{y}| = a_{ud}$, $0 \leq a_{ud} \leq a$. That is, only an a_{ud} sized subset of the attribute dimensions are labeled, and a_{la} other relevant latent dimensions are discovered automatically. The attribute-space is partitioned into observed and latent subspaces: $\mathcal{Y}^a = \mathcal{Y}_{ud}^{a_{ud}} \times \mathcal{Y}_{la}^{a_{la}}$ with $a = a_{ud} + a_{la}$. To support a full spectrum of applications, we should permit $a = a_u$ (traditional attribute learning), and $a = a_l$ (unsupervised latent space).

3.2 Semi-latent Attribute Space Topic Model

LDA. To learn a suitably flexible model for L (Eq. (1)), we generalize LDA[8], modeling each attribute as a topic. LDA provides a generative model for a discrete dataset $D = \{\mathbf{x}_i\}$ in terms of a latent topic y_{ij} for each word x_{ij} given prior topic concentration α and word-topic parameters β. Assuming vector topic proportions α we have

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

where j indexes individual words, $\theta_i|\alpha$ is the Dirichlet topic prior for instance i, $x_{ij}|y_{ij}$ and $y_{ij}|\theta_i$ are discrete with parameters $\beta_{y_{ij}}$ and θ_i.

Variational inference for LDA approximates the intractable posterior $p(\theta_i, \mathbf{y}_i|\mathbf{x}_i, \alpha, \beta)$ in terms of a factored variational distribution: $q(\theta_i, \mathbf{y}_i|\gamma_i, \phi_i) = q(\theta_i|\gamma_i) \prod_j q(y_{ij}|\phi_{ij})$ resulting in the updates:

$$\phi_{ijk} \propto \beta_{x_{ij}k} \exp(\Psi(\gamma_{ik})), \quad \gamma_{ik} = \alpha_{ik} + \sum_j \phi_{ijk}. \tag{3}$$

Semi-Latent Attribute Space (SLAS). With no user defined attributes ($a = a_{la}, a_{ud} = 0$), an a-topic LDA model provides a mapping L from raw data $\mathbf{x}$ to an a-dimensional latent space by way of the variational posterior $q(\theta|\gamma)$. 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 size a^{ud} binary vectors $\mathbf{v}_z^{ud}$ specifying which are present in class z [2,6] We cannot use $\mathbf{v}$ to directly determine or constrain the LDA topic vector $\mathbf{y}^{ud}$. This is because LDA associates each word x_{ij} with a topic y_{ij}, and we don't know word-attribute correspondence. We only know whether each attribute is present in each instance. To enforce this type of constraint, we define a *per instance* prior $\alpha_i = [\alpha_i^{ud}, \alpha_i^{la}]$, setting $\alpha_{i,k}^{ud} = 0$ whenever $v_{z(i),k}^{ud} = 0$. That is, enforcing that instances i of class z lacking an attribute k can never use that attribute the explain the data; but otherwise leaving the inference algorithm to infer attribute proportions and word correspondence. Interestingly, in contrast to other methods, this allows our approach to reason about how strongly each attribute is exhibited in each instance instead of only modeling binary presence and absence.

To learn the latent portion of the attribute-space, we could simply leave the remaining portion α^{la} of the prior unconstrained; however while resulting latent topics/attributes will explain the data, they are not necessarily discriminative. Instead, inspired by [9,7], we split the prior into two components $\alpha_i^{la} = [\alpha_i^{cc}, \alpha^{bn}]$. The first, $\alpha_i^{cc} = \{\alpha_{i,z}\}_{z=1}^{N_z}$, is a series of "class conditional" subsets $a_{i,z}$ corresponding to classes z. For an instance i with label z_i, all the other components $\alpha_{i,z\neq z_i}^{cc}$ are constrained to zero. 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 second component of the latent space prior, α^{bg} is left unconstrained, meaning that in contrast to the CC topics, these "background" topics are shared between all classes. When learned jointly with the CC topics, BN topics are therefore likely to represent common non-discriminative background information [9,7] and thus should be ignored for classification. This is supported by our experiments where we show that better CC topics are learned when BN topics are present.

Classification. Defining the mapping L in Eq. (2) as the posterior statistic γ in SLAS (Eq. (3)), the remaining component to define is the attribute-class mapping S. Importantly, for our complex data, this mapping is not deterministic and 1:1 as is often assumed [2,6]. Like [13], we therefore use standard classifiers to learn this mapping from the γ_is obtained from our SLAS attribute learner.

Zero-Shot Learning (ZSL) with Latent Attributes. To recognize novel classes $\mathcal{Z}^*$, we define the mapping S manually. Existing attribute-learning approaches [2,6] define a simple deterministic prototype $\mathbf{v}_{z^*}^{ud} \in \mathcal{Y}^u$ for class z^*, and classify by NN matching of data to prototype templates. For realistic unstructured video data, huge intra-class variability means that a single prototype is a very poor model of a class, so zero-shot classification will be poor. Counter-intuitively, but significantly more interestingly, *we can actually leverage the latent portion of the attribute-space even without training data for novel class* z^* (so long as there is at least one UD attribute, $a^u \geq 1$) with the following self-training algorithm:

1. Infer attributes γ^* for novel test data X^* (Eq. (3))
2. NN matching in the user-defined space $\gamma^{ud,*}$ against prototypes $\mathbf{v}_{z^*}^{ud}$
3. For each novel class z^*:
   (a) Find top-K most confident test-set matches $\{\gamma_{l,z^*}\}_{l=1}^{K}$
   (b) Self train a new prototype in the full attribute-space: $\mathbf{v}_{z^*} = \frac{1}{K}\sum_l \gamma_{l,z^*}$.
4. NN matching in the full attribute space of γ^* against prototypes $\mathbf{v}_{z^*}$.

Previous ZSL studies are constrained to UD 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 even a loose manual definition of a novel class. We will show later this approach can significantly improve zero-shot learning performance.

4 Experiments

In this section we first introduce our new dataset, and then describe the quantitative results obtained for four types of problems: Multi-task classification; learning with label noise; N-shot learning and ZSL. For each reported experiment, 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.[3].

SVM-UD+LR: SVM attribute classifiers learn available UD attributes. A logistic regression (LR) classifier then learns classes given the probability mapped attribute classifier outputs.[4] This is the obvious generalisation of Direct Attribute Prediction (DAP) [2] to non-deterministic attributes.

SLAS+LR: Our SLAS is learned, then a LR classifier learns classes based on the UD and CC topic profile.

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

[4] LR was chosen over SVM because it is more robust to sparse data.

For all experiments, we cross-validate the regularisation parameters for SVM and LR. For all SVM models, we use the χ^2 kernel. For SLAS, in each experiment, we keep the complexity fixed at 85 topics, up to 69 of which are UD attributes, and the others equally divided between CC and BN latent attributes. The UD part of the SLAS topic profile is estimating the same thing as the SVM attribute classifiers, however the latter are slightly more reliable due to being discriminatively optimised. As input to LR, we therefore actually use the SVM attribute classifier outputs in conjunction with the latent part of our topic profile.

4.1 Unstructured Social Activity Attribute (USAA) Dataset: Classes and Attributes

A new benchmark attribute dataset for social activity video classification and annotation is introduced. We manually annotate the groundtruth attributes for 8 semantic class videos of CCV dataset [16], and select 100 videos per-class for training and testing respectively. These classes were selected as the most complex social group activities. As shown in Fig. 1, a wide variety of attributes have been annotated. The 69 attributes can be broken down into five broad classes: actions, objects, scenes, sounds, and camera movement. We tried our best to exhaustively define every conceivable attribute for this dataset, to make a benchmark for unstructured social video classification and annotation. Of course, real-world video will not contain such extensive tagging. However, this exhaustive annotation gives the freedom to hold out various subsets and learn on the others in order to quantify the effect of annotation density and biases on a given algorithm. These eight classes are birthday party, graduation party, music performance, non-music performance, parade, wedding ceremony, wedding dance and wedding reception (shown in Fig. 3). Each class has a strict semantic definition in the CCV video ontology. Directly using the ground-truth attributes (average annotation density 11 attributes per video) as input to a SVM, the videos can be classified with 86.9% accuracy. This illustrates the challenge of this data: while the attributes are informative, there is sufficient intra-class variability in the attribute-space, that even perfect knowledge of the attributes in an instance is insufficient for perfect classification. The SIFT, STIP and MFCC features for all these videos are extracted according to [16], and included in the dataset. We report the baseline accuracy of SVM-attribute classifiers learned on the whole test set in Fig. 4. Clearly some can be detected almost perfectly, and others cannot be detected given the available features.

4.2 Multi-task Learning

The main advantage of attribute-centric learning when all classes are known in advance is exploiting feature sharing [25]. The statistical strength of data supporting a given attribute can be aggregated across its occurrences in all classes. This treats classification like a multi-task learning problem where the class models share parameters, rather than each class being modelled independently.

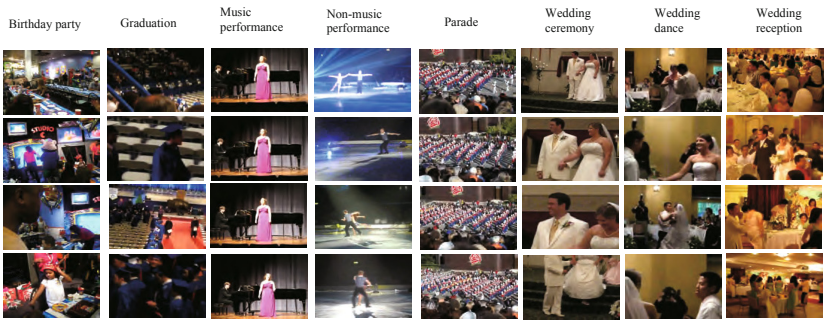


Fig. 3. Example frames from the eight class unstructured social activity dataset

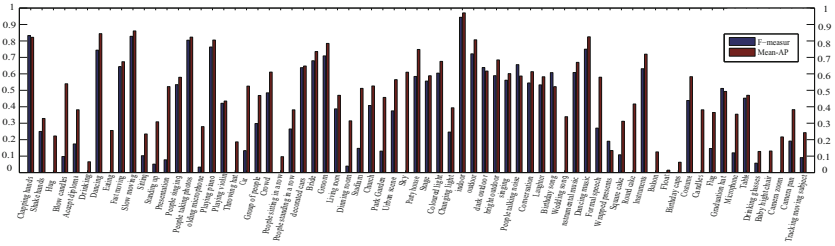


Fig. 4. Attribute-classification accuracy using SVM

Table 1 summarises our results. We first consider the simplest classification scenario where the data is plentiful and the attributes are exhaustively defined. In this case all the models perform similarly. Next, we consider a sparse data variant, with only 10 instances per class to learn from. Here Direct KNN performs poorly due to insufficient data. The attribute models perform better due to leveraging statistical strength across the classes. To the most realistic case of a sparsely defined attribute space, we next limit the attributes to a randomly selected seven every trial, rather than the exhaustively defined 69. In this challenging case SVM+LR performance drops 10% while our SLAS continues to perform similarly, now outperforming the others by a large margin. It is able to share statistical strength among attributes (unlike Direct KNN) and able to fill out the partially-defined attribute space with latent attributes (unlike SVM+LR). Finally, the other challenge in learning from real-world sources of unstructured social video is that the attribute annotations are likely to be very noisy. To

Table 1. Multi-task classification performance (%). (8 classes, chance = 12.5%).

	Direct	SVM+LR	SLAS+LR
100 Inst, 69 UD	**66**	65	65
10 Inst, 69 UD	29	37	**40**
10 Inst, 7 UD	29	27	**36**
10 Inst, 7 UD, attribute noise	27	23	**36**

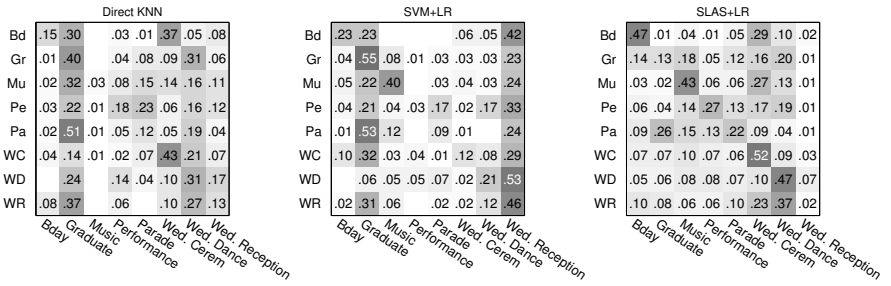


Fig. 5. Confusion matrices for multi-task classification with 10 instances per class

simulate this, we repeated the previous experiment, but randomly changed 50% of attribute bits on 50% of the training videos (so 25% wrong attribute annotations). In this case, performance of the traditional attribute-learning approach is further reduced, while the that of our model is unchanged. This is because our model learns and leverages a whole space of latent attributes to produce a robust representation which can compensate for noise in the UD attributes.

Fig. 5 shows the confusion matrices for the 10 instance, 7 attribute task. The matricies for the traditional Direct KNN and SVM attribute classification have vertical bands indicating consistent misclassifications. Our SLAS has the clearest diagonal structure with little banding, indicating no consistent errors.

4.3 N-Shot Transfer Learning

In transfer learning, one assumes ample examples of a set of source classes, and sparse examples of a *disjoint* set of target classes. To test this scenario, in each trial we randomly split our 8 classes into two disjoint groups of four source and target classes. We use all the data from the source task to train our attribute learning models (SLAS and SVM), and then use these to obtain the attribute profiles of the target task. Using the target task attribute profiles we perform N-shot learning, with the results summarised by Table 6. Importantly, traditional attribute learning approaches cannot deal with zero attribute situations. Our SLAS performs comparably or better than both Direct-KNN and SVM+LR for zero, seven and 34 attributes.This illustrates the robustness of our model to the density of the attribute-space definition. Importantly, standard attribute-learning (SVM+LR) cannot function with zero attributes, but our attribute model maintains a significant margin over Direct KNN in this case.

4.4 Zero-Shot Learning

One of the most interesting capabilities of attribute-learning approaches is zero-shot learning. Like N-shot learning, the task is to learn transferrable attribute knowledge from a source dataset for use on a disjoint target dataset. However, no training examples of the target are available. Instead, user manually specifies the definition of each novel class in the semantic attribute space. Zero-shot learning

	1-shot			5-shot		
	Direct KNN	SVM+LR	SLAS+LR	Direct KNN	SVM+LR	SLAS+LR
0 UD	30	-	**34**	34	-	**42**
7 UD	30	32	**33**	34	43	**44**
34 UD	30	**37**	35	34	47	**48**

Fig. 6. N-shot classification performance (%). (4 classes, chance = 25%)

is often evaluated in simple situations where classes have unique 1:1 definitions in the attribute-space [2]. For our unstructured social data, strong intra-class variability violates this assumption, making evaluation slightly more subtle. We compare two approaches: "continuous" prototypes, where a novel class definition is given by continuous values in attribute-space, and "binary" prototypes, where the novel class is defined as a binary attribute vector. These correspond to two models of human provided semantic knowledge: continuous or thresholded probability that a new class has a particular attribute. E.g., saying that cakes and candles are definite attributes of a birthday party vs saying they might occur with 90% and 80% probability respectively. To simulate these two processes of prior knowledge generation, we take the mean and the thresholded mean (as in [13,10]) of the attribute profiles for each instance.

Our results are summarised in Table 2. Using latent attributes to support the user-defined attributes (Sec. 3.2) allows our SLAS model to improve on the conventional user-defined attribute only approach to zero-shot learning. Interestingly, continuous definition of class prototypes is a significantly more powerful approach for both methods (Table 2, Continuous vs Binary). To illustrate the value of our other contribution, we also show the performance of our model when learned without free background topics (SLAS (NF)). The latent attribute approach is still able to improve on using pure user-defined attribute, but by a smaller margin. The BN topics generally improve performance by segmenting the less discriminative dimensions of the latent attribute space and allowing them to be ignored by the classifier.

Table 2. Zero-shot classification performance (%). (4 classes, chance = 25%).

Continuous			Binary		
UD	UD+Latent		UD	UD+Latent	
SVM-DAP	SLAS	SLAS (NF)	SVM-DAP	SLAS	SLAS (NF)
38	**45**	41	31	**36**	31

5 Conclusions

Summary. In this paper we have considered attribute learning for the challenging task of understanding unstructured multi-party social activity video. To promote study of this topical issue, we introduced a new multi-modal dataset with extensive detailed annotations. In this context, a serious practical issue is

the limited availability of annotation relative to the number and complexity of relevant concept classes. We introduced a novel semi-latent attribute-learning technique which is able to: (i) flexibly learn a full semantic-attribute space when attribute space is exhaustively defined, or completely unavailable, available in a small subspace (i.e., present but sparse), or available but noisy; (ii) perform conventional and N-shot while leveraging latent attributes and (iii) go significantly beyond existing zero-shot learning approaches (which only use defined attributes), in leveraging latent attributes.. In contrast, standard approaches of direct classification or regular attribute-learning fall down in some portion of the contexts above (Section 4).

Future Work. There are a variety of important related open questions for future study. Thus far, our attribute-learner does not consider inter-attribute correlation explicitly (like most other attribute learners with the exception of [13]). This can be addressed relatively straightforwardly by generalising the correlated topic model (CTM) [26] for our task instead of regular LDA [8]. A correlated attribute model should produce commensurate gains in performance to those observed elsewhere [13].

We have made no explicit model [27] of the different modalities of observations in our data. However explicit exploitation of the different statistics and noise-processes of the different modalities is an important potential source of improved performance and future study (e.g., learning modality-attribute correlations and inter-modality correlations via attributes).

The complexity of our model was fixed to a reasonable value throughout (i.e., the size of the semi-latent attribute/topic-space), and we focused on learning with attribute-constraints on some sub-set of the topics. More desirable would be a non-parametric framework which could infer the appropriate dimensionality of the latent attribute-space automatically. Moreover, we ware able to broadly separate foreground and "background" topics via the different constraints imposed; however it is not guaranteed that background topics are irrelevant, so not using them in classification may be sub-optimal. A more systematic way (e.g., [7]) to automatically segment discriminative "foreground" and distracting "background" attributes would be desirable.

References

1. Wang, C., Blei, D., Li, F.F.: Simultaneous image classification and annotation. In: Proc. CVPR (2009)
2. Lampert, C.H., Nickisch, H., Harmeling, S.: Learning to detect unseen object classes by between-class attribute transfer. In: CVPR, pp. 951–958 (2009)
3. Farhadi, A., Endres, I., Hoiem, D., Forsyth, D.: Describing objects by their attributes. In: Proc. CVPR (2009)
4. Farhadi, A., Endres, I., Hoiem, D.: Attribute-centric recognition for cross-category generalization. In: CVPR (2010)
5. Parikh, D., Grauman, K.: Relative attributes. In: Proc. ICCV (2011)
6. Palatucci, M., Hinton, G., Pomerleau, D., Mitchell, T.M.: Zero-shot learning with semantic output codes. In: Proc. NIPS (2009)

7. Hospedales, T., Gong, S., Xiang, T.: Learning tags from unsegmented videos of multiple human actions. In: Proc. ICDM (2011)
8. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent dirichlet allocation. Journal of Machine Learning Research 3, 993–1022 (2003)
9. Hospedales, T., Li, J., Gong, S., Xiang, T.: Identifying rare and subtle behaviours: A weakly supervised joint topic model. PAMI (2011)
10. Farhadi, A., Endres, I., Hoiem, D., Forsyth, D.: Describing objects by their attributes. In: Proc. CVPR (2009)
11. Mahajan, D., Sellamanickam, S., Nair, V.: A joint learning framework for attribute models and object descriptions. In: Proc. ICCV (2011)
12. Parikh, D., Grauman, K.: Interactively building a discriminative vocabulary of nameable attributes. In: CVPR (2011)
13. Liu, J., Kuipers, B., Savarese, S.: Recognizing human actions by attributes. In: Proc. CVPR (2011)
14. Wang, Y., Mori, G.: Human action recognition by semilatent topic models. TPAMI 31, 1762–1774 (2009)
15. Niebles, J.C., Wang, H., Fei-Fei, L.: Unsupervised learning of human action categories using spatial-temporal words. IJCV 79, 299–318 (2008)
16. Jiang, Y.G., Ye, G., Chang, S.F., Ellis, D., Loui, A.C.: Consumer video understanding: A benchmark database and an evaluation of human and machine performance. In: ICMR (2011)
17. Yanagawa, A., Loui, E.C., Luo, J., Chang, S.F., Ellis, D., Jiang, W., Kennedy, L.: Kodak consumer video benchmark data set: concept definition and annotation. In: Proc. ACM MIR (2007)
18. Gupta, A., Srinivasan, P., Shi, J., Davis, L.S.: Understanding videos, constructing plots learning a visually grounded storyline model from annotated videos (2009)
19. Wang, M., Hua, X.S., Hong, R., Tang, J., Qi, G.J., Song, Y.: Unified video annotation via multigraph learning. IEEE Trans. Cir. and Sys. for Video Technol. (2009)
20. Tang, J., Yan, S., Hong, R., Qi, G.J., Chua, T.S.: Inferring semantic concepts from community-contributed images and noisy tags. In: Proc. ACM MM (2009)
21. Tang, J., Hua, X.S., Qi, G.J., Song, Y., Wu, X.: Video annotation based on kernel linear neighborhood propagation. IEEE Transactions on Multimedia (2008)
22. Snoek, C.G.M., Worring, M.: Concept-based video retrieval. Foundations and Trends in Information Retrieval 4, 215–322 (2009)
23. Snoek, C.G.M., Huurnink, B., Hollink, L., de Rijke, M., Schreiber, G., Worring, M.: Adding semantics to detectors for video retrieval. IEEE Transactions on Multimedia 9, 975–986 (2007)
24. Dollar, P., Rabaud, V., Cottrell, G., Belongie, S.: Behavior recognition via sparse spatio-temporal features. In: VS-PETS, pp. 65–72 (2005)
25. Salakhutdinov, R., Torralba, A., Tenenbaum, J.: Learning to share visual appearance for multiclass object detecti. In: Proc. CVPR (2011)
26. Blei, D., Lafferty, J.: A correlated topic model of science. Annals of Applied Statistics 1, 17–35 (2007)
27. Putthividhy, D., Attias, H.T., Nagarajan, S.S.: Topic regression multi-modal latent dirichlet allocation for image annotation. In: Proc. CVPR, pp. 3408–3415 (2010)