Learning Tags from Unsegmented Videos of Multiple Human Actions

Timothy M. Hospedales, Shaogang Gong and Tao Xiang
Queen Mary University of London, UK, E1 4NS
{tmh,sgg,txiang}@eecs.qmul.ac.uk

***Abstract*—Providing methods to support semantic interaction with growing volumes of video data is an increasingly important challenge for data mining. To this end, there has been some success in recognition of simple objects and actions in video; however most of this work requires strongly supervised training data. The supervision cost of these approaches therefore renders them economically non-scalable for real world applications.**

In this paper we address the problem of learning to annotate and retrieve semantic tags of human actions in realistic video data with sparsely provided tags of semantically salient activities. This is challenging because of (1) the multi-label nature of the learning problem and (2) realistic videos are often dominated by (semantically uninteresting) background activity un-supported by any tags of interest, leading to a strong irrelevant data problem.

To address these challenges, we introduce a new topic model based approach to video tag annotation. Our model simultaneously learns a low dimensional representation of the video data, which dimensions are semantically relevant (supported by tags), and how to annotate videos with tags. Experimental evaluation on three different video action/activity datasets demonstrate the challenge of this problem, and value of our contribution.

***Index Terms*—action recognition; annotation; tag learning; topic model;**

I. Introduction

Managing and exploiting growing volumes of video data is an important challenge for data mining and computer vision research. Semantic video interactions such as content based indexing, search, retrieval, classification and annotation are therefore all topical problems. One of the most useful content classes for semantic interaction is that of human actions, with applications in security, retailing, entertainment and sports. Towards this end, the problem of human activity analysis has recently been studied intensively [1]. Most of these contemporary approaches to learning human activity models require pre-segmented or strongly labeled data for training [30], [18], [29], [7]. Specifically, the user must specify three dimensional bounding cuboids in (x,y,t) for each training example of an interesting activity. This is a rather imposing barrier to real world use. It would be much more useful if a system could learn from the sparse semantic tags (text) that might already have been associated with a video, for example in the metadata management system of a video sharing site, social network, or home media center. Ideally, such a system would leverage a large existing collection of video data and associated existing tags to learn a model of the dynamic visual content relevant to each tag. Once such a model is learned, newly added video can be annotated with likely tags without further effort from the user and these annotated tags can then be used in support of indexing, retrieval, etc.

This video tag learning problem can be viewed as a multi-label learning problem [24] in that each instance (video) may be associated with multiple labels. There has been progress on multi-label learning with static image data [6], [26], [16]. However solving the same problem with video data is harder for two reasons: (i) the video instances are now unbounded in (temporal) size, potentially making the multi-label aspect of the problem harder; and (ii) the visual support of the tags associated with a video may be a fairly small percentage of the overall volume of the video. That is, most of the moving pixels may be semantically uninteresting, thereby posing a strong irrelevant data problem [12]. Fig. 1 illustrates this challenge in two sample video frames. Here, more visual interest points are related to irrelevant background activity (crosses) than the punching and waving actions of interest (dots). In fact, the problem can be worse than visualized as the temporal extent of a video could include an unbounded number of frames where the tagged actor has not yet arrived, or has already left.

In this paper, we present a framework for joint generative topic modeling, relevance determination, and annotation of video, which we call VTT (Video Tags and Topics). The general idea is to learn a topic model which jointly predicts the visual words (bags of visual feature interest points) and associated semantic tags (text/phrases). That is, the model learns a low-dimensional topic decomposition of the video database simultaneously with a built-in generalized linear model to predict the tags based on the topic profile. Performing these tasks jointly helps the model to learn topics which are useful in discriminating actions rather than merely providing a good generative model of the video. Additionally, we generalize the notion of joint generative modeling of data and tags via topics to include tags which depend on an unknown subset of the topics. The partitioning of the topic space into background dimensions – which explain away visual data unsupported by tags – and salient dimensions – which are predictive of tags – is learned automatically. This feature enables us to cope with data where visual evidence corresponding to the tags may include only a small portion of the total video volume.

A. Related work

There is now a fairly extensive literature [1] on video activity recognition and detection including approaches based on techniques such as dynamic Bayesian networks (DBNs)

Figure 1. Example video frames in which actions of interest are visually weak compared to ongoing background activity. Dots denote detected interest points which may be relevant (dots) or irrelevant (crosses) to the punching or waving actions.

[28], support vector machines (SVMs) [23], branch and bound [30], [7], random forests / hough transforms [29], constellation models [9], [20] and topic models [21]. We do not review activity recognition in detail here except to point out that most approaches requires at least segmented or strongly labeled training data. Specifically visual features corresponding to a single activity label are assumed to dominate each training instance (video clip). That is, training video must be chosen or temporally segmented such that each clip contains exactly one interesting action; and chosen or spatially segmented such that background features are minimized. This assumption is unrealistic for most videos of human activities in a unconstrained social environments.[1] To our knowledge this study is the first attempt to learn automatic association of semantic tags of human activities from multi-label and non-segmented video data.

In other domains such static image annotation [6], there has been recent success in learning to predict multiple tags from each instance. This is variously known as annotation, tagging, attribute learning or multi-label learning [24], [6]. Topic models [26], [16], [22] based on CorrLDA [4] have been shown to be suited for this task, typically by learning a multinomial distribution over tags conditional on each topic. However, this approach limits annotation accuracy because (i) using a single multinomial output at test time predicts tags competitively [8] and (ii) being additive in their predictions, topics can not provide negative information about a particular anti-correlated action. In general, image annotation models are not suitable for addressing our problem because human activity tags are sparse, and there is more uninteresting background activity to deal with, i.e. not all topics learned can be attributed to human actions of interest in video.

Other approaches include multiple Bernoulli models [8] and SVMs [6], [19]. One issue [24] for annotation models is whether tags are predicted independently [8], a simpler task but ignoring any correlation between tags, or jointly [32], a harder task but exploiting more information. In our model, although each tag is predicted independently for computational efficiency and robustness, inter-tag correlation is still exploited because predictions are made based on the shared low dimensional topic space.

Our work builds on the class of models known as topic models (such as latent Dirichlet allocation (LDA) [5]), which provide a generative model for discrete data in terms of a lower dimensional mixture of latent topics. Recent LDA variants have included supervised response variables [3], [14], [31]. However, these have thus far been restricted to single label tasks. CorrLDA [4] adds a simple mechanism for multi-label annotation to LDA, and we will compare against it explicitly. None of these models deals with the strong irrelevant data problem posed by video data studied here.

One interesting debate in the literature is the difference, if any, between "tags" and "classes", and what that implies for how they should each be modeled. Typically the class of an instance summarizes the entire instance while tags refer to a subset of the instance (e.g., in image domain an image of class snow-boarding may have tags snow, mountain, sky, trees, person) [26]. For this reason tags are typically modeled more simply (e.g., topic conditional multinomial [26], [16], [22]), while classes have more complex models (e.g., the root of a Bayesian Network [16], or topic conditional classifier [26]). In our case the actions corresponding to tags are potentially quite complex (similarly to [30], [29], [7], [28], [21]). However, the unbounded temporal extent of video means many can also occur in a single instance. That is, although we are formally solving an annotation problem, our tags have the complexity of full classes as considered by previous action recognition work ([30], [29], [7], [28], [21]). In contrast to existing annotation models [4], [26], [16], we therefore learn a full generalized linear model to predict tags (activities) from topics.

Most learning methods are challenged by large amounts of irrelevant data. Standard supervised classification typically overcomes the problem with a separate feature selection step [12], although it may also be solved generatively by modeling both the relevant and irrelevant data [13]. Existing multi-label/annotation methods assume that tags are dense enough that most data is supported by a tag [4], [26], [16], [22], [32], [19], [8]. Solving a multi-label task in an irrelevant data context is particularly challenging because each pixel needs to be disambiguated between both association with each known class and an arbitrarily complex background process, which can be more dominant. Similarly to [13], our approach is to model *all* the data via topics, while learning which *subset* of the topics is predictive of the tags. However [13] only addresses standard single label learning, and requires a pre-specified partition of topics into relevant and background categories. In contrast, we address multi-label learning and our relevance learning flexibly allocates topics in any proportion to a foreground or background partition according to the data.

II. Methods

A. Preprocessing and Representation

Topic models requires a discrete bag of visual words representation of video. To extract suitable and informative features, we apply the space-time interest point detector [15] to discover

[1] We note that one apparently similar prior action recognition study [9] learns from "unsegmented" video, but for the single label case, and in the simplistic Weizmann dataset [2] without background activity. It is therefore effectively segmented in our context.

interest points in each video $j = 1 \dots N_v$. After applying k-means vector quantization to the interest point descriptors, each video in the dataset $\mathbf{X} = \{\mathbf{x}_j\}_{j=1}^{N_v}$ is represented by a bag of N_j discrete codewords $\mathbf{x}_j = \{x_{ij}\}_{i=1}^{N_j}$. We use a codebook size of $N_w = 2000$.

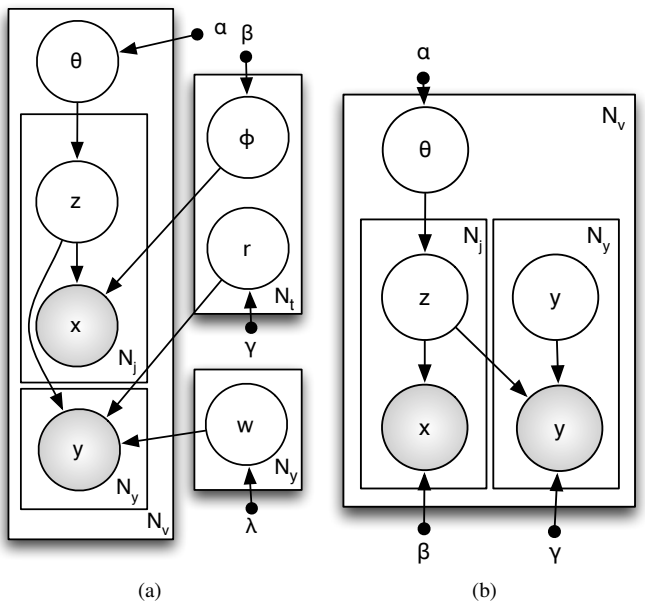
B. VTT Generative Model

In this section we describe our VTT model (Fig. 2(a)) in detail. In VTT each video j contains a bag of visual words $\{x_{ij}\}_{i=1}^{N_j}$ and tags $\{y_{lj}\}_{l=1}^{N_y}$ to be modeled (Fig. 2, shaded nodes). All N_y tag variables are considered to be binary and conditionally independent given their latent parents. Visual words and tags are correlated between and across modalities by virtue of being generated via the lower dimensional space of shared latent topics $\mathbf{z}_j$. Each topic $k = 1 \dots N_t$ represents a component of a visible action in the video database, and may be shared between actions with related structure or allocated uniquely to an action as suggested by the data. The tags are assumed to be relevant to an unknown subset of these topics which is to be discovered.

The generative process is as follows: The activity distribution in video j is uniquely defined by Dirichlet topic proportions θ_j. From these proportions, the topic z_{ij} for each visual word (i, j) is independently generated by a multinomial distribution $z_{ij} \sim p(z_{ij}|\theta_j)$. The visual word itself is finally generated from the multinomial of it's associated topic, $x_{ij} \sim p(x_{ij}|\Phi, z_{ij})$. The separation of topics corresponding to salient activities supported by tags, and non-salient background activities is generated by sampling independently a binary mask variable $r_k \sim p(r_k|\gamma)$ for each topic k. Finally, based on the latent topics and their relevance, each tag y_{lj} is generated independently according to a Bernoulli distribution $y_{lj} \sim p(y_{lj}|\mathbf{z}_j, \mathbf{r}, \mathbf{W})$. Learning the parameters of this distribution will identify the relation between topics and tags. The complete generative model is specified by:

$$
\begin{aligned}
p(\boldsymbol{\phi}_k|\beta) &= \text{Dir}(\boldsymbol{\phi}_k; \beta), \\
p(\boldsymbol{\theta}_j|\boldsymbol{\alpha}) &= \text{Dir}(\boldsymbol{\theta}_j; \boldsymbol{\alpha}), \\
p(z_{ij}|\boldsymbol{\theta}_j) &= \text{Multi}(z_{ij}; \boldsymbol{\theta}_j), \\
p(x_{ij}|z_{ij}, \boldsymbol{\Phi}) &= \text{Multi}(x_{ij}; \boldsymbol{\phi}_{z_{ij}}), \\
p(r_k|\gamma) &= \text{Bern}(r_k|\gamma), \\
p(\mathbf{w}_l|\lambda) &= \mathcal{N}(\mathbf{w}_l|0, \lambda), \\
p(y_{lj}|\mathbf{r}, \mathbf{z}_j, \mathbf{W}) &= \text{Bern}(y_{lj}|\rho_{lj}), \\
\rho_{lj} &= 1/\left(1 + \exp(-\mathbf{w}_l^T \frac{\overline{\mathbf{z}} \otimes \mathbf{r}}{\sum_k \overline{z}_k r_k})\right),
\end{aligned}
$$

where the N_t length deterministic vector $\overline{\mathbf{z}}$ represents the empirical topic proportions in $\mathbf{z}_j$ and $\overline{\mathbf{z}} \otimes \mathbf{r}$ indicates the selection of dimensions in $\overline{\mathbf{z}}$ according to true elements of binary vector $\mathbf{r}$. The full joint distribution of observed $O = \{\mathbf{x}_j, \mathbf{y}_j\}_{j=1}^{N_v}$ and latent variables $\{H = \{\mathbf{z}_j, \theta_j\}_{j=1}^{N_v}, \{\phi_k, r_k\}_{k=1}^{N_t}, \{\mathbf{w}_l\}_{l=1}^{N_y}\}$ given parameters $P = \{\boldsymbol{\alpha}, \beta, \gamma, \lambda\}$ of our model is:



(a) (b)

Figure 2. Graphical model representation of (a) our VTT model and (b) CorrLDA [4]. Shaded nodes are observed for training. For testing y is unobserved.

$$
\begin{aligned}
p(O, H|P) &= \prod_k p(\boldsymbol{\phi}_k|\beta) p(r_k|\gamma) \prod_j p(\boldsymbol{\theta}_j|\boldsymbol{\alpha}) \\
&\cdot \prod_i p(z_{ij}|\boldsymbol{\theta}_j) p(x_{ij}|z_{ij}, \boldsymbol{\phi}_{z_{ij}}) \prod_l p(y_{lj}|\mathbf{r}, \mathbf{z}_j, \mathbf{w}_l) p(\mathbf{w}_l|\lambda). \quad (1)
\end{aligned}
$$

Parameters $\boldsymbol{\alpha}$ and β represent our prior belief about the proportions of the topics in the dataset, and the sparsity of visual words within each topic. $\mathbf{w}_l$ represents the mapping between latent topics and tag l. γ represents our prior belief about the proportion of topics in the dataset which are salient (supported by tags).

To train our model, we provide a corpus of processed videos $\mathbf{X}$ and associated tags $\mathbf{Y}$. Learning the hidden variables and parameters from the training set corresponds to: discovering a lower dimensional representation of the video and tags (vectors $\mathbf{z}_j$ and $\boldsymbol{\theta}_j$), learning to predict each tag y in the vocabulary (vectors $\mathbf{w}_l$), and learning which aspects the video representation are salient for tagging (vector $\mathbf{r}$). Once learned, we test our model by providing a new video $\mathbf{x}^*$ without associated tags, for which the model infers the latent topics $\mathbf{z}^*$, and predicts the new tags $\mathbf{y}^*$.

C. Model Learning, Inference and Testing

1) Inference: Exact probabilistic learning in VTT is intractable, so in this section we derive a stochastic expectation maximization (EM) algorithm based on Gibbs sampling and iterative conditional modes (ICM) [10] for approximate learning in our model. As for standard LDA [5], we analytically integrate out the conjugate-prior Dirichlet parameters Φ and Θ. A standard EM approach to learning would then be to alternate inference of latent variables $p(\mathbf{Z}, \mathbf{W}, \mathbf{r}|\mathbf{X}, \mathbf{Y}, \boldsymbol{\alpha}, \beta, \gamma, \lambda)$ with parameter estimation

$$\{\boldsymbol{\alpha}, \beta, \gamma, \lambda\} \leftarrow \text{argmax} \sum_{\mathbf{Z},\mathbf{W},\mathbf{r}} p(\mathbf{Z}, \mathbf{W}, \mathbf{r}|\mathbf{X}, \mathbf{Y}, \boldsymbol{\alpha}, \beta, \gamma, \lambda) \cdot \ln p(\mathbf{X}, \mathbf{Y}, \mathbf{Z}, \mathbf{W}, \mathbf{r}|\boldsymbol{\alpha}, \beta, \gamma, \lambda)$$

In our case, we resort to Gibbs sampling and ICM to estimate the posterior $p(\mathbf{Z}, \mathbf{W}, \mathbf{r}|\mathbf{X}, \mathbf{Y}, \boldsymbol{\alpha}, \beta, \gamma, \lambda)$. For Gibbs sampling, we need the posterior of each hidden variable conditioned on all the other variables. It will first be useful to denote the likelihood $\mathcal{L}_j = p(\mathbf{y}_j|\mathbf{z}_j, \mathbf{r}, \mathbf{W})$ of a document's tag set $\{y_{lj}\}_{l=1}^{N_y}$ as:

$$\mathcal{L}_j = \prod_{l=1}^{N_y} \text{Bern}(y_{lj}; (1 + \exp(-\mathbf{w}_l^T \frac{\overline{\mathbf{z}} \otimes \mathbf{r}}{\sum_k \overline{z}_k r_k}))^{-1}). \quad (2)$$

The update for each latent topic z_{ij} is then

$$\begin{aligned} p(z_{ij}|\mathbf{Z}_{-ij}, \mathbf{r}, \mathbf{W}, \mathbf{X}, \mathbf{Y}) & \\ \propto\ & p(\mathbf{y}_j|\mathbf{z}_j, \mathbf{r}, \mathbf{W}) p(x_{ij}|\mathbf{Z}, \mathbf{X}_{-ij}) p(z_{ij}|\mathbf{Z}_{-ij}), \\ =\ & \mathcal{L}_j \frac{n_{xz}^{-ij} + \beta}{\sum_x n_{xz}^{-ij} + \beta} \frac{n_{zj}^{-ij} + \alpha_z}{\sum_z n_{zj}^{-ij} + \alpha_z}, \end{aligned} \quad (3)$$

where we have dropped conditioning on the parameters for clarity. The update (3) contains three terms which reflect: the predictive fit of the topics to the tags, the fit to the visual words, and to the topic prior. Here $-ij$ means "excluding item (i, j)". n_{xz}^{-ij} indicates the counts of topic z being associated with word x excluding (i, j) and n_{yj}^{-ij} indicates the counts of topic z in document j.

Since the dimension of the weight vector $\mathbf{w}$ depends on the relevance variable $\mathbf{r}$, they must be updated together. We have

$$\begin{aligned} p(r_k, \mathbf{W}|\mathbf{r}_{-k}, \mathbf{Y}, \mathbf{Z}) \ &\propto\ p(\mathbf{Y}|\mathbf{Z}, \mathbf{W}, \mathbf{r}) p(r_k|\mathbf{r}_{-k}) p(\mathbf{W}), \\ &=\ \prod_{j=1}^{N_v} \mathcal{L}_j \gamma \prod_l^{N_y} \exp -\lambda \mathbf{w}_l^T \mathbf{w}_l. \end{aligned} \quad (4)$$

For efficiency, we take a hybrid Gibbs and ICM approach, sampling the relevance variable r_k:

$$r_k \sim \frac{1}{K} \prod_{j=1}^{N_v} \mathcal{L}_j \gamma, \quad (5)$$

and updating $\mathbf{W}$ to its MAP value. To maximize $\mathbf{W}$ we obtain the gradient of (4) with respect to each parameter vector $\mathbf{w}_l$:

$$\nabla_{\mathbf{w}_l} \ln p(\mathbf{Y}|\mathbf{Z}, \mathbf{r}, \mathbf{W}) p(\mathbf{W}) = \sum_j (y_{lj} - \rho_{lj}) \overline{\mathbf{z}}_j - 2\lambda \mathbf{w}_l. \quad (6)$$

$\mathbf{w}_l$ is then optimized by a fast L-BGFS gradient based optimizer [17]. We note that with this approach, while the updates (5) and (6) for $\mathbf{r}$ and $\mathbf{W}$ are costly, they are only performed $\mathcal{O}(N_t)$ times per Gibbs sweep. The $\mathcal{O}(N_j)$ per Gibbs sweep topic updates (3) are the dominant computational cost for VTT as for most topic models.

2) Learning: Model parameters $\{\boldsymbol{\alpha}, \beta, \gamma\}$ are updated by Gibbs-EM (i.e, EM updates using the approximate posterior obtained by averaging the latent variable samples obtained over a certain lag) resulting in the same updates as in [13], [25]. The weight prior (regularization) parameter λ is also updated periodically by internal 3-fold cross-validation. The number of topics, N_y is the only manually set parameter in our model. However, we emphasize that it is not important to tune, because good results can always obtained by setting a large number of topics and allowing hyperparameter learning of vector $\boldsymbol{\alpha}$ to reduce the weight of unused topics [25]. At the end of the training phase, point estimates of the word-topic parameters $\hat{\Phi}$ are computed as the mean of their Dirichlet posteriors $\hat{\boldsymbol{\phi}}_k = \frac{n_{xk}+\beta}{\sum_x n_{xk}+\beta}$ [11]. The current tag-topic weights $\hat{\mathbf{W}}$, relevance variables $\hat{\mathbf{r}}$ and topic hyperparameters $\hat{\boldsymbol{\alpha}}$ are also stored. We simulate a batch of N_s independent Markov chains and store these parameters for the last sample of the chain. We observed this approach to obtain statistically independent samples to be more efficient than taking repeated samples from the same chain, while also being more amenable to parallelism.

3) Testing: To use the learned model to annotate a new video $\mathbf{x}^*$, the predictive tag distribution $p(\mathbf{y}^*|\mathbf{x}^*, \mathbf{X}, \mathbf{Y}) = \sum_{\mathbf{z}^*} p(\mathbf{y}^*|\mathbf{z}^*, \mathbf{X}, \mathbf{Y}) p(\mathbf{z}^*|\mathbf{x}^*, \mathbf{X}, \mathbf{Y})$ is of required. To approximate this, we again simulate a batch of $s = 1 \ldots N_s$ Markov chains, each initialized with the corresponding parameter vector $\{\hat{\boldsymbol{\alpha}}, \hat{\mathbf{W}}, \hat{\mathbf{r}}, \hat{\Phi}\}_s$ obtained from training. Only (3) is iterated to infer the topic posterior for the test document $p(\mathbf{z}^*|\mathbf{x}^*, \{\hat{\boldsymbol{\alpha}}, \hat{\mathbf{W}}, \hat{\mathbf{r}}, \hat{\Phi}\}_s)$. Samples of the test document topic profiles $\mathbf{z}_s^*$ drawn from these distributions are used to approximate the final tag distribution for the test document as

$$p(\mathbf{y}^*|\mathbf{x}^*, \mathbf{X}, \mathbf{Y}) \approx \sum_s p(\mathbf{y}^*|\mathbf{z}_s^*, \hat{\mathbf{W}}_s). \quad (7)$$

III. Experiments

A. Illustrative Example

First, let us apply our proposed VTT model to a simulated dataset. This serves three purposes: to illustrate the mechanism of our model; to validate its correct behaviour on data which is non-trivial but with known ground-truth; and to provide insight into its properties compared to other models as a function of background noise which we can control precisely here. The experiment is illustrated in Fig. 3. To generate training data, we defined nine tags and associated visual patterns (Fig. 3(a)) and six irrelevant background patterns (Fig. 3(b)). For each instance, two tags were randomly selected and their patterns used to as priors to generate half of the words for the instance, and the remaining half of the words were generated from the background patterns in random proportions. The resulting dataset is illustrated in Fig. 3(c). Although this is synthetic data, it is still visually quite challenging to distinguish any structure in Fig. 3(c). We generated 200 such images with 25 words each and associated lists of relevant tags, and trained a VTT model with $N_y = 16$ topics. The model learned an appropriate latent representation and associated partitioning: a set of twelve relevant topics – axis aligned bars in various

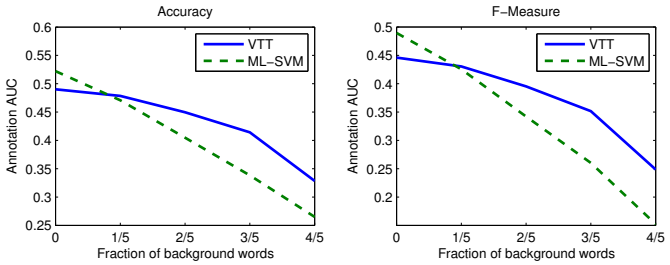


Figure 4. VTT model performance as a function of the fraction of words generated from the irrelevant background distribution.

positions (Fig. 3(d)), which can be composed to express the visual patterns of each tags; and a set of four irrelevant topics – various diagonal bars (Fig. 3(e)), which explain the remaining non-tagged background components of the data. Note that partitioning of the topics and the proportions of the partitioning (Fig. 3(d) vs (e)) were learned automatically. Despite the noise in each sample (Fig. 3(c)), reconstructing the implicit internal pattern learned by the model for each tag by visualizing $\mathbf{\Phi w}_l$ (Fig. 3(f)) shows a fair match to the original definitions (Fig. 3(a)). Note that in this dataset each relevant topic (Fig. 3(d)) is shared between *three* tag patterns (Fig. 3(f)), and it is a particular combination of topics – as learned by $\mathbf{w}_l$ – that defines a tag. Fig. 3(g) illustrates the learned model being used to correctly label a test image with its relevant tags.

Finally, we quantified the performance of our model at annotating test images with the nine tags as a function of the fraction of words generated by irrelevant patterns (Fig. 3(b)). The results (Fig. 4) verify that performance decays much more slowly for VTT than the standard multi-label SVM approach ([6], Sec. III-C) to learning annotations. This property of our model will be important to obtain good performance with the real world video data studied next.

B. Video Datasets

In this study we experiment with three video datasets of increasing challenge level[2]. **KTH:** The KTH action dataset [23] is aimed at action classification and exhibits one person performing a single action in each video. It is fairly clean and simple, however due to its relatively large size (600 videos of 25 people performing six action categories) and extensive prior study, we consider it here. To render the dataset suitable for multi-label learning, we concatenate pairs and triples of randomly chosen videos and tag them with the labels of the action categories of the component clips. This results in 291 two-action videos or 196 three action videos. **MSR:** The MSR dataset [30] is aimed at action detection, and exhibits multiple people performing three actions in more crowded and busy scenes. This dataset is interesting because of the realistic and challenging levels of background noise, but limited in only exhibiting three actions in total (punching, waving and clapping). In its original form each video clip has the full set of three actions, which is unsuitable for any multi-label learning approach. We therefore convert the dataset for annotation learning by randomly cropping the videos in time such that only two salient actions occur in each video. This results in 102 two-action videos. **CPSM:** The CPSM "sports minute" dataset is a new dataset collected by us from youtube containing two years worth news highlights about college sports. There are 74 videos in total with 10 different sports (Table. I). This is a real world dataset, so the proportions of each activity and number of labels per clip are variable.

C. Experimental Conditions

We compare the performance of our model against two existing approaches: multi-label SVM [6] and CorrLDA [4].

- ML-SVM: A simple approach to multi-label classification is to decompose the problem into N_y independent binary single-label SVM problems of learning to separate instances with and without each tag. This simplistic approach ignores the correlation between tags, but it often performs well in practice [6]. We took care to obtain the best performance by optimizing the SVM kernel choice and hyper-parameters by cross-validation, and compensating for imbalanced data by appropriate asymmetric weighting of the cost parameter.
- CorrLDA: Correspondence LDA (Fig. 2(b)) learns a set of topic-conditional multinomial distributions for generating tags. This idea has been used successfully by various image annotation studies: [4], [16], [26]. The important assumption here is that a tag corresponds to (is generated by) the topic of a single visual word. This is in contrast to VTT where the entire profile of topics is potentially used to determine the tag.

To quantify the performance of the models, we take two approaches: an annotation approach, in which the accuracy of predicted annotations for test videos is evaluated; and an information retrieval approach, in which test videos relevant to a specified tag are retrieved, and the relevance of the retrieved videos is evaluated. For annotation, we consider two standard measures: hamming accuracy and the F-measure[3] between the estimated and true tag list. These measures can be generated from annotations obtained from ranking (asking the model to return the top-N) most likely tags – most useful if there are a fixed number of tags per video; or by detecting tags surpassing a certain probability threshold which is varied obtain an accuracy-threshold curve and F-measure-threshold curve. For information retrieval, we use the F-measure evaluated as a function of the number of videos retrieved.

We train each model on 2/3s of the data and test on the remaining 1/3s of the data. For VTT, we use $N_y = 32$ topics. To train VTT, we simulate 3 Markov chains with 100 iterations of burn in, followed by 100 iterations where hyper-parameters are updated every 5 iterations. For testing, we run 100 iterations before drawing the samples $\mathbf{z}_s^*$ used for prediction in (7).

[2]To be made available at http://www.eecs.qmul.ac.uk/~tmh/datasets/

[3]Defined as $F = \frac{2 \times Precision \times Recall}{Precision + Recall}$ with a range of 0 to 1.

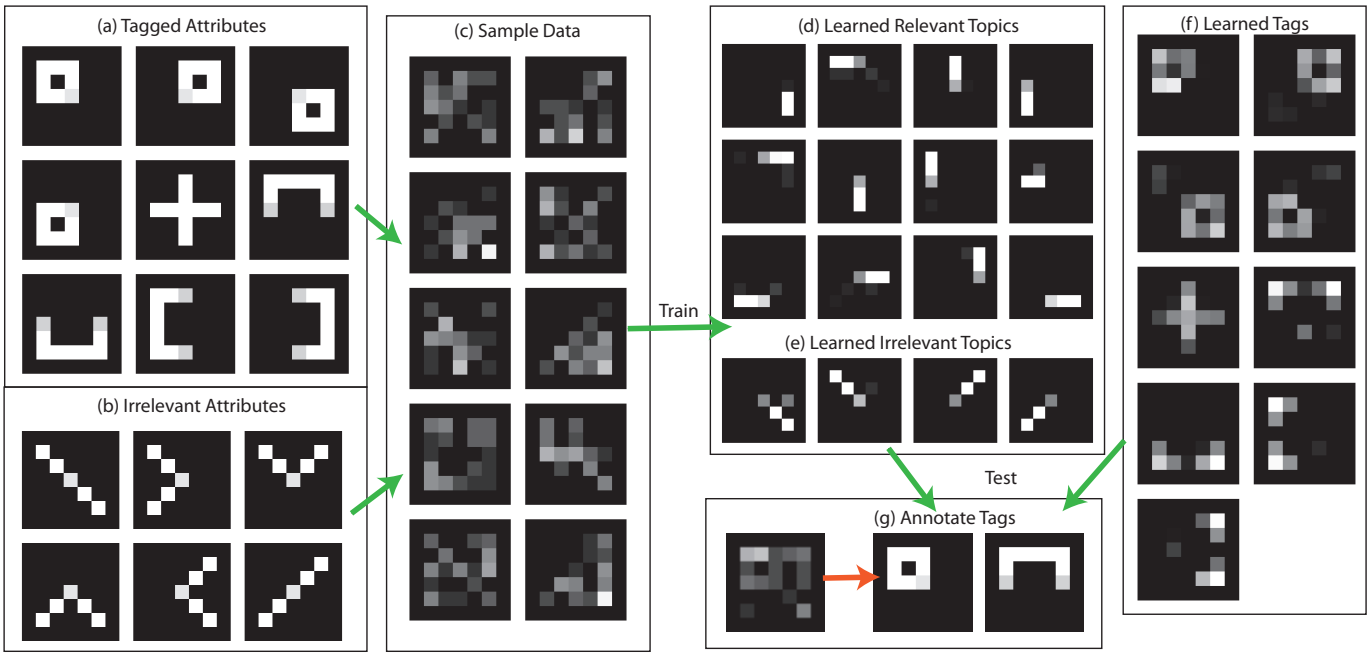


Figure 3. Illustration of how our VTT model works via a synthetic data example. Prototype (a) relevant and (b) irrelevant attributes are used to generate (c) sample data. The learned latent representation induces a partition into (d) relevant and (e) irrelevant topics. (f) Internal representation of the learned salient attributes. (g) Correctly annotating a test image with tags.

	Activities										Labels per Clip					
	Swim	Football	Soccer	Tennis	Volleyball	Running	Baseball	Golf	Basketball	Wrestling	1	2	3	4	5	6
Clips	6	37	27	14	26	8	33	4	36	28	4	21	35	13	2	1

Table I
CPSM DATASET DETAILS

D. Results

1) KTH: Table II shows the average test performance on the KTH dataset for each model rounded to the nearest percentage point. The AUC rows indicate scores obtained by varying the tag detection threshold and evaluating the area under the F-measure curve. The Rank-N rows indicate scores obtained by assigning the top-N most likely tags. Note that Rank 2 chance is 55% for accuracy and 33% for F-measure.

For the easier case of two tags per video, ML-SVM slightly outperforms VTT. However, for the harder case of three tags per video where the multi-label ambiguity is greater, VTT performs best by a larger margin. We note that for the KTH dataset, single label learning and classification performance reported in the literature [27] is around 90% when using the same features and SVM classifier used here. The slightly lower scores here reflect the more challenging nature of the multi-label learning problem and the more stringent AUC evaluation criterion. CorrLDA [4] is generally the worst for each experiment. It performs relatively well at ranking tags (Rank-N rows), but very poorly at actually detecting them reliably (AUC rows). This is expected because CorrLDA's single multinomial output predicts tags competitively, which disadvantages it at tag detection.

		VTT		ML-SVM		CorrLDA	
Tags/Video	Eval	FM	Acc	FM	Acc	FM	Acc
2	A-Rank2	81	85	**84**	**87**	73	79
2	A-AUC	61	75	**71**	**81**	23	67
3	A-Rank3	**76**	**76**	63	63	70	70
3	A-AUC	**59**	**64**	47	57	16	52

Table II
KTH DATASET ANNOTATION PERFORMANCE. RESULTS IN TERMS OF AVERAGE F-MEASURE AND HAMMING ACCURACY.

2) MSR: The average test performance on the MSR dataset for each model is shown in Table III. We note that the performance of all models is much worse than for the KTH dataset. For the tag ranking task, performance is not far above chance baseline (55% for accuracy, 66% for F-measure). This reflects two key challenge factors: the reduced amount of training data available for this dataset, and the huge amount of irrelevant background activity in this dataset (Fig. 1). By evaluating the ground truth detection cuboids provided by [30], we observed that only 5-50% of the visual words in each video are even within the bounding box of the target activities. This joint uncertainty in both relevance to the problem, and which words are related to which tag render this task very challenging. Our

	VTT		ML-SVM		CorrLDA	
Eval	FM	Acc	FM	Acc	FM	Acc
A-Rank2	**69**	**65**	59	54	68	63
A-AUC	**63**	**59**	57	55	30	45
IR-AUC	**60**	-	50	-	57	-

Table III
MSR DATASET ANNOTATION PERFORMANCE. RESULTS FOR ANNOTATION AND RETRIEVAL IN TERMS OF AVERAGE F-MEASURE AND HAMMING ACCURACY.

	VTT		ML-SVM		CorrLDA	
Eval	FM	Acc	FM	Acc	FM	Acc
A-AUC	**60**	**80**	57	76	17	72
A-Rank2	**83**	**84**	83	84	81	83
IR-AUC	48	-	47	-	**50**	-

Table IV
CPSM ANNOTATION PERFORMANCE. RESULTS FOR ANNOTATION AND RETRIEVAL IN TERMS OF AVERAGE F-MEASURE AND HAMMING ACCURACY.

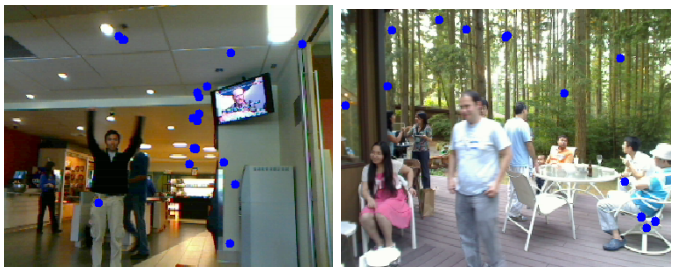


Figure 5. Examples of background topics learned in the MSR dataset. Illustration by plotting visual words in the sample frame assigned to the background topic.

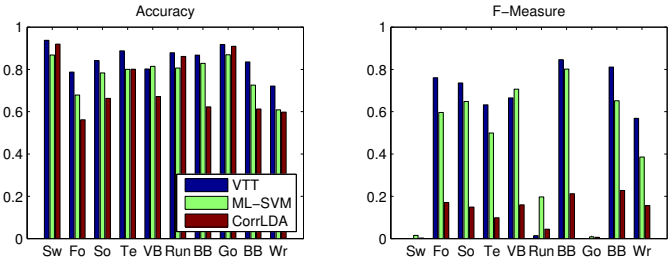


Figure 6. CPSM Annotation Performance. AUC breakdown by activity class.

model is better able to deal with this challenging data due to its automatic partitioning of background and foreground topics which help to explain away the irrelevant data. Fig. 5 illustrates this for the MSR dataset, showing interest points allocated to two background topics, which roughly correspond to explaining the image-plane movement of indoor lighting sources and edges in outdoor forests in response to camera motion. These effects are uncorrelated to interesting activities and so are allocated correctly by the model to background topics. We also evaluated an information retrieval task, (IR-AUC row) for which VTT also performs best.

3) CPSM: Finally, quantitative performance for the most realistic and challenging CPSM dataset is shown in Table IV, and a breakdown by individual activities is given in Fig. 6. For VTT and ML-SVM models, annotation performance is good for seven of the ten classes, and very poor for swimming, running and golf. This is largely due to the fact that the latter three types are the activities with the least representation in the dataset (see Table I), so evidently there was not enough data to learn a good model in these cases. Again, CorrLDA performs competitively at ranking and information retrieval (Table IV, A-Rank2 and IR-AUC rows), but fails at actually reliably annotating individual clips (Table IV, A-AUC). This is important because actual annotation rather than ranking is arguably the more relevant measure of performance for many important applications (e.g., automatic tagging on video sharing sites or home media systems).

To demonstrate the challenging nature of this dataset and the output that our model is capable of producing, Fig. 7 illustrates a set of three videos with perfectly successful annotations and two additional videos with a false positive and a false negative. Note that un-tagged but common elements such as news readers are explained away by background topics in VTT.

We note that in terms of the learned mapping $\mathbf{W}$ between topics and annotation tags, individual topics can be shared between tags or allocated uniquely to a specific tag. To illustrate this, Fig. 8(a) shows a learned foreground topic broadly representing "running people at medium viewing distance" which has been discovered and is shared (significant positive weight in multiple $\mathbf{w}_l$) between soccer, American football and running activities. In contrast, Fig. 8(b) illustrates a learned foreground topic specifically representing the stereotyped motion of a baseball pitcher, which has been allocated solely to the baseball activity (only positive weight in $\mathbf{w}_l$ for baseball).

IV. DISCUSSION

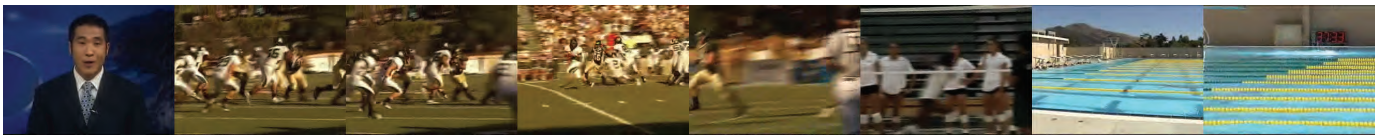
We have introduced a powerful solution to the topical problem of learning to annotate realistic videos with tags representing human activities. This approach generalizes existing supervised topic models [3], [4] to address the challenging problem of multi-label learning in the presence of background noise. Learning the latent topics jointly with a tag annotation model induces more discriminative representations to be formed than in a purely unsupervised topic model of video data. Topics can be shared between multiple activities or allocated specifically to single activities as required. In contrast to additive annotation models such as [4], [16], [26], topics in VTT can also provide specific negative information about particular tags with which they are unlikely to be associated.

Relevance detection is modeled internally in the latent space of the VTT model. This is in contrast to traditional supervised learning where feature selection is often treated as an independent and suboptimal wrapper process around a black box learner [12]. Performing feature selection in the latent space is also much more computationally efficient than the conventional approach of searching for a good subset of dimensions in the (much larger) original input space. In summary, our model searches directly and jointly for a low

Tags: American Football, Soccer, Volleyball.

Tags: Basketball, Wrestling

Tags: Tennis, Baseball



Tags: American Football, Volleyball, Swimming

Tags: Wrestling, Baseball, (Basketball)

Figure 7. Example annotation results for CPSM dataset. Each row shows a news video. Three perfectly successful annotations and two containing errors are shown.

dimensional representation of videos and tags; a partitioning of relevant vs irrelevant latent topics and a predictive mapping for tag annotation. Positive results on three datasets of increasing challenge level support the value of this contribution.

REFERENCES

[1] J. K. Aggarwal and M. S. Ryoo. Human activity analysis: A review. *ACM Computing Surveys*, to appear.

[2] M. Blank, L. Gorelick, E. Shechtman, M. Irani, and R. Basri. Actions as space-time shapes. In *IEEE International Conference on Computer Vision*, volume 2, pages 1395–1402, 2005.

[3] David Blei and Jon McAuliffe. Supervised topic models. In *Neural Information Processing Systems*, 2007.

[4] David M. Blei and Michael 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] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.

[6] Matthew R. Boutell, Jiebo Luo, Xipeng Shen, and Christopher M. Brown. Learning multi-label scene classification. *Pattern Recognition*, 37(9):1757 – 1771, 2004.

[7] Liangliang Cao, Zicheng Liu, and T. S. Huang. Cross-dataset action detection. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1998–2005, 2010.

[8] S. L. Feng, R. Manmatha, and V. Lavrenko. Multiple bernoulli relevance models for image and video annotation. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2004.

[9] R. Filipovych and E. Ribeiro. Learning human motion models from unsegmented videos. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2008.

[10] Brendan J. Frey and Nebojsa Jojic. A comparison of algorithms for inference and learning in probabilistic graphical models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(9):1392 – 1416, September 2005.

[11] Thomas L. Griffiths and Mark Steyvers. Finding scientific topics. *Proceedings of the National Academy of Sciences*, 101:5228–5235, 2004.

[12] Isabelle Guyon and Andre Elisseeff. An introduction to variable and feature selection. *Journal of Machine Learning Research*, 3:1157–1182, 2003.

[13] 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.

[14] Simon Lacoste-Julien, Fei Sha, and Michael I. Jordan. Disclda: Discriminative learning for dimensionality reduction and classification. In *Neural Information Processing Systems*, 2008.

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

[16] Li-Jia Li, R. Socher, and Li Fei-Fei. Towards total scene understanding: Classification, annotation and segmentation in an automatic framework. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2036–2043, 2009.

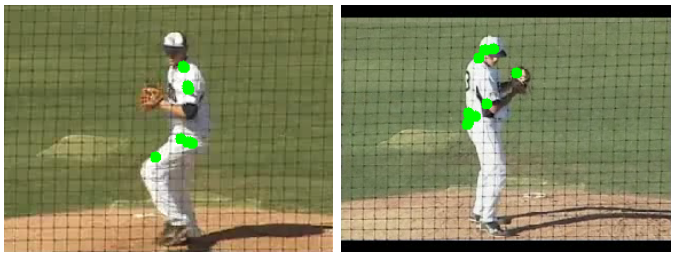
[17] D. Liu and J. Nocedal. On the limited memory method for large scale optimization. *Mathematical Programming B*, 45(3):503–528, 1989.

[18] K. Mikolajczyk and H. Uemura. Action recognition with motion-appearance vocabulary forest. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2008.

[19] Nam Nguyen. A new svm approach to multi-instance multi-label learning. In *International Conference on Data Mining*, pages 384–392, 2010.

[20] J. C. Niebles and Li Fei-Fei. A hierarchical model of shape and appearance for human action classification. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2007.

(a) Running people topic



(b) Pitching topic

Figure 8. Examples of learned topics which are (a) shared and (b) uniquely allocated.

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

[22] D. Putthividhy, H. T. Attias, and S. S. Nagarajan. Topic regression multi-modal latent dirichlet allocation for image annotation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3408–3415, 2010.

[23] C. Schuldt, I. Laptev, and B. Caputo. Recognizing human actions: a local svm approach. In *International Conference on Pattern Recognition*, volume 3, pages 32–36, 2004.

[24] G. Tsoumakas, I. Katakis, and I. Vlahavas. *Mining Multi-label Data*, chapter Data Mining and Knowledge Discovery Handbook. Springer, 2 edition, 2010.

[25] Hanna M. Wallach, David Mimno, and Andrew McCallum. Rethinking lda: Why priors matter. In *Neural Information Processing Systems*, 2009.

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

[27] Heng Wang, Muhammad Muneeb Ullah, Alexander Klaser, Ivan Laptev, and Cordelia Schmid. Evaluation of local spatio-temporal features for action recognition. In *British Machine Vision Conference*, 2009.

[28] Tao Xiang and Shaogang Gong. Beyond tracking: Modelling activity and understanding behaviour. *International Journal of Computer Vision*, 61(1):21–51, 2006.

[29] A. Yao, J. Gall, and L. Van Gool. A hough transform-based voting framework for action recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2061–2068, 2010.

[30] J. Yuan, Z. Liu, and Y. Wu. Discriminative video pattern search for efficient action detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, (99), 2011. Early Access.

[31] Jun Zhu, Amr Ahmed, and Eric P. Xing. Medlda: maximum margin supervised topic models for regression and classification. In *International Conference on Machine Learning*, 2009.

[32] Shenghuo Zhu, Xiang Ji, Wei Xu, and Yihong Gong. Multi-labelled classification using maximum entropy method. In *Proceedings of the International ACM SIGIR conference on Research and Development in Information Retrieval*, 2005.