Noname manuscript No.
(will be inserted by the editor)

Learning Relevance Restricted Boltzmann Machine for Unstructured Group Activity and Event Understanding

Fang Zhao · Yongzhen Huang · Liang Wang · Tao Xiang · Tieniu Tan

Received: date / Accepted: date

Abstract Analyzing unstructured group activities and events in uncontrolled web videos is a challenging task due to 1) the semantic gap between class labels and low-level visual features, 2) the demanding computational cost given high-dimensional low-level feature vectors and 3) the lack of labeled training data. These difficulties can be overcome by learning a meaningful and compact mid-level video representation. To this end, in this paper a novel supervised probabilistic graphical model termed relevance Restricted Boltzmann Machine (ReRBM) is developed to learn a low-dimensional latent semantic representation for complex activities and events. Our model is a variant of the Restricted Boltzmann Machine (RBM) with a number of critical extensions: (1) sparse Bayesian learning is incorporated into the RBM to learn features which are relevant to video classes, i.e., discriminative; (2) binary stochastic hidden units in the RBM are replaced by rectified linear units in order to better explain complex video contents and make variational inference tractable for the proposed model; and (3) an efficient variational EM algorithm is formulated for model parameter estimation and inference. We conduct extensive experiments on two recent challenging benchmarks: the Unstructured Social Activity Attribute dataset and the Event Video dataset. Experimental results demonstrate that the relevant features learned by our model provide better semantic and discriminative description for videos than a number of alternative supervised latent variable models, and achieves state of the art performance in terms of classification accuracy and retrieval precision, particularly when only a few labeled training samples are available.

Keywords Representation learning · Video analysis · Restricted Boltzmann Machine · Sparse Bayesian learning

F. Zhao (✉) · Y. Huang · L. Wang · T. Tan
Center for Research on Intelligent Perception and Computing,
Institute of Automation, Chinese Academy of Sciences, Beijing, China
E-mail: fang.zhao@nlpr.ia.ac.cn

Y. Huang
E-mail: yzhuang@nlpr.ia.ac.cn

L. Wang
E-mail: wangliang@nlpr.ia.ac.cn

T. Xiang
School of Electronic Engineering and Computer Science,
Queen Mary, University of London, London, United Kingdom
E-mail: t.xiang@qmul.ac.uk

T. Tan
E-mail: tnt@nlpr.ia.ac.cn

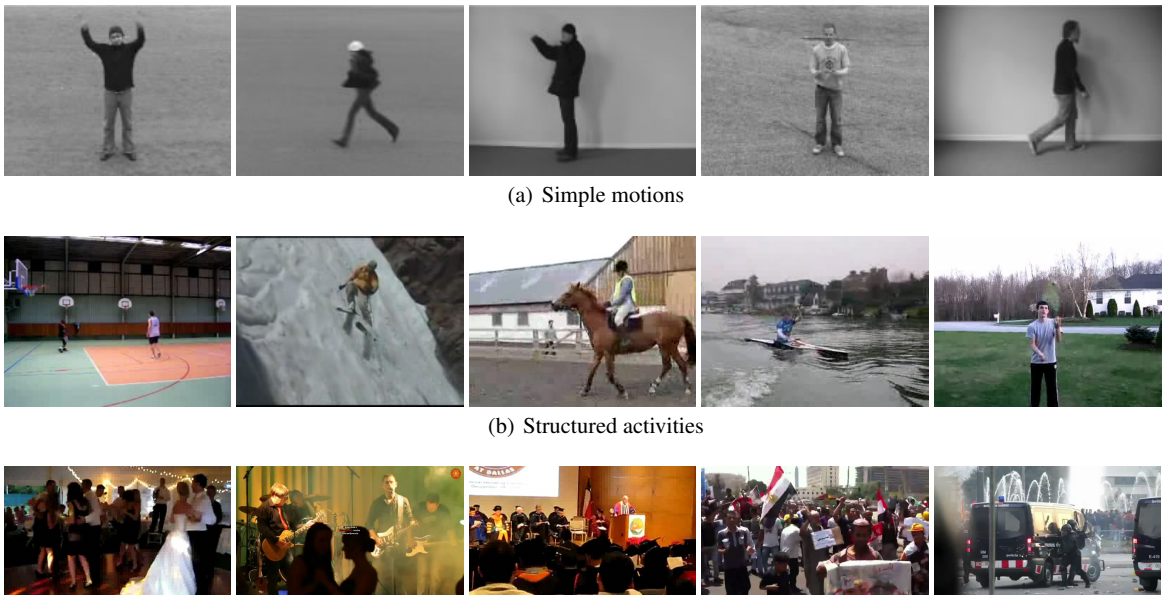
1 Introduction

Every minute, 100 hours of videos are uploaded to YouTube – equivalent to 16 years of new content every day.[1] In 2013, Web videos account for 53% of internet downstream traffic in North America, with YouTube alone around 19%.[2] This growing volume of data demands effective and efficient ways for users to organize, browse and search videos, and for video-sharing website operators to provide accurate personalised recommendations and sensibly targeted advertisements. Commercial search engines rely on text metadata associated with videos, including the title, description or tags provided by users; but these metadata are typically sparse, incomplete, noisy and sometimes inconsistent with the video content. As a result, automatic video analysis techniques, such as classification, retrieval and recommendation, have received increasing interests.

Among the great variety of videos uploaded on the internet, the videos containing unstructured group activities (e.g., wedding reception and graduation ceremony) and events

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

[2] http://www.hollywoodreporter.com/news/video-accounts-53-percent-internet-655203



(a) Simple motions

(b) Structured activities

(c) Unstructured group activities and events

Fig. 1 Comparison of sample frames from different types of video datasets.

(e.g., anti-government demonstrations and riots) pose the ultimate challenge for automated video content analysis (see Figure 1 (c)). These videos often feature a large number of people and objects interacting with each other. Importantly different type of videos may share many common characteristics, e.g. wedding reception and graduation ceremony all have large crowds gathering which act and sound similarly, e.g. clapping hands. Videos containing unstructured activities and events differ significantly from the videos targeted by most existing techniques such as those in the KTH (Schuldt et al, 2004), UCFSports (Rodriguez et al, 2008) and UCF50 (Reddy and Shah, 2013) datasets. The videos in these datasets are short clips containing simple motions or structured activities without complex multi-object interaction (see Figure 1 (a) and (b)). In a conventional approach designed for understanding the contents of these simpler videos, the following steps are taken. First, a set of training samples are labeled; second, low-level features are extracted from the videos to form a representation; and finally, these features are fed into a model for clustering, classification or retrieval. However, this standard pipeline is not suitable for analysing unstructured group activities or events. This is due to the following problems: 1) The semantic gap between class labels and low-level visual features – complex videos contain rich semantic concepts that cannot be represented directly and explicitly by the low-level features. 2) The demanding computational cost given high-dimensional low-level feature vectors – to represent complex video content, features of high dimensionality (e.g. thousands) are typically extracted. This results in complex models with a large number of model parameters and intractable learning and inference algorithms for large scale problems. 3) The lack of labeled training data – although there are literally unlimited videos available online, few of them are adequately labelled. Manually annotating large quantities of videos is often infeasible. On the other hand, freely available tags associated with the videos are often sparse and noisy, sometimes even irrelevant to the video content.

To overcome these problems, recently semantic concept (or attribute) based mid-level representations have been proposed to bridge the semantic gap (Wei et al, 2011; Fu et al, 2012) and provide a compact representation. However, those methods need human defined ontology and manual annotation of attribute vectors for each class or instance, and thus they scale poorly to large scale problems. Critically, attribute based approaches (e.g. Fu et al, 2012) are principally designed for transfer learning to recognize unseen classes without any training data, i.e., zero-shot learning. For tasks such as recognizing seen classes with labeled training samples, attributes have not yet proven convincingly as an effective alternative to low-level feature based representations.

In this paper, we aim to learn rather than handcraft a mid-level representation of unstructured group activities and events, in order to bridge the semantic gap and reduce the dimensionality of the visual representation. Critically, these learned mid-level features need to be discriminative, that is, relevant to the video classes, so as to facilitate different tasks such as classification and retrieval. Such a mid-level feature can be learned as latent variables in a probabilistic graphical model. Latent feature representation learned by probabilistic graphical models have been widely used to analyze text, images and videos. Examples of such models include

topic models (Zhu et al, 2012; Rasiwasia and Vasconcelos, 2013; Wang and Mori, 2009) and Restricted Boltzmann Machines (Salakhutdinov and Hinton, 2009; Larochelle et al, 2012; Taylor et al, 2010). The representations in these models are constructed purely from data without the need for human intervention, thus are more scalable. However, how to learn more discriminative and compact latent representations from complex video content remains an unsolved problem (see Section 1.1). In particular, unstructured group activities and events captured by amateurs using hand-held camcorders or mobile phones often contain some distracting visual patterns such as camera jitters and background movements and noise. A useful mid-level representation must be able to filter out these patterns and keep the more meaningful ones that can be used to discriminate different activity and event classes.

To this end, a novel hierarchal probabilistic graphical model is proposed to discover compact and discriminative/relevant mid-level representations for unstructured group activities and events in videos. The model, termed relevance Restricted Boltzmann Machine (ReRBM), is based on the Restricted Boltzmann Machine (RBM) (Smolensky, 1986) which is an undirected graphical model with a bipartite structure. Compared to the standard RBM, our ReRBM has a built-in relevance measure based on sparse Bayesian learning (Tipping, 2001) to make the learned mid-level features more discriminative and compact. In addition, binary stochastic hidden units in the RBM are replaced by rectified linear units (Nair and Hinton, 2010), which allows each unit to express more information for better explaining video data containing complex content and also makes variational inference tractable for the proposed model. By employing a simple quadratic bound on the log-sum-exp function (Bohning, 1992), an efficient variational EM algorithm is developed for parameter estimation and inference. Furthermore, our model can be easily extended to accommodate multimodal feature inputs (e.g. visual and audio) necessary for modeling complex video contents.

1.1 Related work

Low-level features for video representation – The problem of extracting video features has been extensively studied on standard datasets ranging from the simplest KTH to the more realistic UCF50. One of the earliest works on designing video low-level features are (Laptev, 2005; Laptev et al, 2008) which proposed to detect space-time interest points and aggregated their descriptors into a compact representation based on bag-of-words. Wang et al (2011) investigated dense trajectories based representation for videos. Recently, Gopalan (2013) proposed a joint sparsity-based representation by decomposing a video sequence into that observed by spatially/temporally distributed receivers. Till now, most of the existing works (Turaga et al, 2008) were focused on controlled and well-structured videos containing limited contents (e.g., clean background and little camera motions). They follow the standard pipeline, i.e., firstly designing and extracting low-level features and then leaning classifiers. However this pipeline is unsuitable for understanding unstructured group activities and events due to the semantic gap problem mentioned earlier. Different from these approaches, our method uses the low-level feature representation as model input and learns both a mid-level feature representation and a classifier in a single model.

Semantic attributes for video representation – Compared with simple videos, unstructured group activity and event analysis in uncontrolled videos has been much less explored. As mentioned earlier, these complex videos pose a number of significant challenges that are beyond the capabilities of most existing approaches using low-level feature representations. In order to address these challenges, recently semantic concepts (or attributes) have been studied as a mid-level representation, which are originally proposed for static images (Lampert et al, 2009, 2013; Farhadi et al, 2009), and then extended for videos (Liu et al, 2011; Wei et al, 2011). Yang and Shah (2012) attempted to learn data-driven concepts from multi-modality video data in an unsupervised manner. Izadinia and Shah (2012) considered modelling co-occurrence relations among the low-level events in a graph to detect complex events, which require extra labeling for the low-level events. Most relevant to our work is a recent work that learns video attributes to analyze unstructured group activities (Fu et al, 2012), wherein a semi-latent attribute space was introduced, consisting of human-defined attributes, class-conditional and background latent attributes. Besides, an extended Latent Dirichlet Allocation (LDA) (Blei et al, 2003) was formulated to model those attributes as latent topics. Different from (Fu et al, 2012), our approach is weakly supervised and automatically discovers a set of discriminative latent feature representations without human annotated attributes. In addition, our approach differs significantly from these semantic attribute based approaches in that (1) we do not require human defined ontologies, and (2) mid-level representation and classifier are learned jointly in a single model.

Probabilistic graphical models – Our relevance Restricted Boltzmann Machine (ReRBM) is one type of probabilistic graphical models (PGMs). PGMs have been employed before for learning mid-level latent feature representations. Most existing models are either unsupervised, or supervised but unable to learn discriminative latent representations (Rasiwasia and Vasconcelos, 2013). The ones that are most closely related to our model are the maximum entropy discrimination LDA (MedLDA) (Zhu et al, 2012) and the supervised Restricted Boltzmann Machines (sRBM) (Larochelle et al, 2012), both of which have been success-

fully applied to document semantic analysis. MedLDA integrates the max-margin learning and hierarchical directed topic models by optimizing a single objective function with a set of expected margin constraints. MedLDA tries to estimate parameters and find latent topics in a max-margin sense, which is different from our model that relies on the principle of automatic relevance determination (Neal, 1995). sRBM also uses class labels to learn some discriminative features. Instead of point estimation of classifier parameters in sRBM, our proposed model learns a sparse posterior distribution over parameters within a Bayesian paradigm. This makes the learned features more compact and discriminative, resulting in better classification and retrieval performance as demonstrated by our experiments (see Section 4).

Sparse Bayesian learning – It is a general probabilistic framework to obtain sparse solutions of parameters in regression and classification tasks. It has been used in the Relevance Vector Machine (Tipping, 2001) to discern basis functions which are relevant to good predictions. Here we use sparse Bayesian learning to select latent features related to classes for semantic representation learning.

Representation learning – Our work is also related to the concept of feature learning or representaiton learning (Bengio et al, 2012). In particular, our model extends Restricted Bolzmann Machines (RBMs) which have been explored recently for representation learning (Ranzato and Hinton, 2010; Sun et al, 2013). Apart from extending RMBs to an undirected and directed hybrid graph model, and introducing sparse Bayesian learning, a key extension is to replace the binary stochastic hidden units in the RBM with real valued ones via rectified linear units (Nair and Hinton, 2010), which allows each unit to express more information for better explaining complex video data. The limitations of a standard RBM caused by its binary visible and hidden units have long been acknowledged and efforts have been injected to generalize it to real value data for the visible units (Hinton and Salakhutdinov, 2006; Ranzato and Hinton, 2010). Nevertheless, no attempt has been made to generalise the binary hidden units to real values so far. Recently RBMs have been extended for deep learning, either by joint training of multiple layers of hidden units as in Deep Boltzmann Machines (Desjardins et al, 2012), or by using them as the final layers on top of a deep convolutional neural network (Hinton et al, 2006; Sun et al, 2013) which take the raw image data as input. Our work is orthogonal to these works and can be integrated into these deep representation learning architectures.

1.2 Contributions

Our main contributions include: 1) We propose a unified framework based on a single hierarchal model to learn both mid-level video representation and classifier jointly. 2) By leveraging labels associated with videos and sparse priors on classifier weights, we extend the standard RBMs to extract meaningful and compact latent features which are more suitable for complex video classification and retrieval. 3) We develop an efficient learning and inference algorithm for the proposed model via variational inference. 4) We conduct extensive experiments to demonstrate that the latent feature representation learned by the proposed model has more discriminative power resulting in better classification and retrieval performance compared with other state of the art alternatives. A preliminary version of the work was reported in (Zhao et al, 2013). In comparison with (Zhao et al, 2013), apart from more comprehensive description, analysis and experiments, this paper formulates a more generalized latent variable model based on the RBM for representation learning, which is not confined to the topic model as discussed in (Zhao et al, 2013).

The rest of this paper is organized as follows. The background knowledge is briefly introduced in Section 2. The proposed model is presented in Section 3. Experimental evaluations are presented in Section 4. Finally, Section 5 concludes this paper and discusses the future work.

2 Background

2.1 Restricted Boltzmann Machine and its variants

A Restricted Boltzmann Machine (RBM) is an undirected graphical model which can be used to learn features unsupervised from input data and has been successfully applied to a variety of represenation learning tasks involving high dimensional data such as images and videos (Hinton and Salakhutdinov, 2006; Ranzato and Hinton, 2010). As shown in Figure 2, the standard RBM has a two-layer architecture, in which the bottom layer represents stochastic visible units $\mathbf{v} \in \{0,1\}^D$ and the top layer represents stochastic hidden units $\mathbf{h} \in \{0,1\}^F$, that is, both sets of variables are binary. The energy function of the state $\{\mathbf{v}, \mathbf{h}\}$ is defined as follows:

$$E(\mathbf{v},\mathbf{h};\theta) = -\sum_{i=1}^{D}\sum_{j=1}^{F} W_{ij} v_i h_j - \sum_{i=1}^{D} a_i v_i - \sum_{j=1}^{F} b_j h_j, \quad (1)$$

where $\theta = \{\mathbf{W}, \mathbf{a}, \mathbf{b}\}$, W_{ij} is the weight connected with v_i and h_j, a_i and b_j are the bias terms of visible and hidden units respectively. The joint distribution over the visible and hidden units is given by:

$$P(\mathbf{v},\mathbf{h};\theta) = \frac{1}{\mathcal{Z}(\theta)} \exp(-E(\mathbf{v},\mathbf{h};\theta)),$$
$$\mathcal{Z}(\theta) = \sum_{\mathbf{v}}\sum_{\mathbf{h}} \exp(-E(\mathbf{v},\mathbf{h};\theta)), \quad (2)$$

where $\mathcal{Z}(\theta)$ is the partition function.

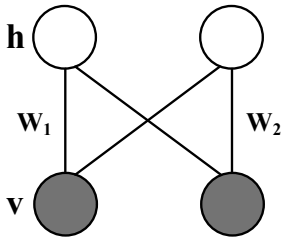


Fig. 2 Restricted Boltzmann Machine: a two-layer undirected graphical model. Visible/observed nodes are shaded.

Some generalizations to binary visible units in the stardard RBM have been proposed to improve the applicability of the model. When the input data are word count vectors, the Replicated Softmax model (Salakhutdinov and Hinton, 2009) (an undirected topic model) can be used for modeling, which is a family of RBMs that share parameters. Let a multinomial visible unit $\mathbf{v} \in \mathbb{N}^N$ represent a word count vector (N is the size of the vocabulary). Then the energy function of the state $\{\mathbf{v}, \mathbf{h}\}$ is defined by:

$$E(\mathbf{v}, \mathbf{h}; \theta) = -\sum_{i=1}^{N}\sum_{j=1}^{F} W_{ij} v_i h_j - \sum_{i=1}^{N} a_i v_i - K \sum_{j=1}^{F} b_j h_j, \tag{3}$$

where $K = \sum_i v_i$ is the total number of words in a document and $\mathbf{h}$ can be seen as latent topics. It has been shown that the Replicated Softmax model outperforms directly probabilistic latent topic models such as Latent Direchlet Allocation (LDA) in terms of both the generalization performance and the retrieval accuracy on text datasets (Salakhutdinov and Hinton, 2009).

The Gaussian RBM (Hinton and Salakhutdinov, 2006) can be used to model real-valued input data. Let $\mathbf{v} \in \mathbb{R}^D$ be real-valued visible units. The energy function is defined by:

$$E(\mathbf{v}, \mathbf{h}; \theta) = \\ -\sum_{i=1}^{D}\sum_{j=1}^{F} \frac{1}{\sigma_i} W_{ij} v_i h_j - \sum_{i=1}^{D} \frac{(v_i - a_i)^2}{2\sigma_i^2} - \sum_{j=1}^{F} b_j h_j, \tag{4}$$

where σ is the standard deviation of the input data. We shall show later that how the descriptive power of the learned features can be enhanced in our model by replacing the binary hidden units with real-valued ones.

Unlike directed graphical models, the conditional distribution of the RBM over hidden units is factorial due to its special bipartite graph structure. Thus the inference of latent variables is straightforward. However, exact maximum likelihood learning in this model is still intractable. The contrastive divergence (Hinton, 2002) approximation is often used to estimate model parameters in practice.

2.2 Low-level video representation

Our model takes low-level feature representations as the model inputs. We consider local keypoint features (such as scale-invariant feature transform (SIFT) (Lowe, 2004), spatial-temporal interest points (STIP) (Laptev, 2005) and mel-frequency cepstral coefficients (MFCC) (Logan, 2000)) which capture static visual appearance, space-time visual appearance and audio features respectively. Then we encode them into fixed-dimensional representation vectors. The three types of low-level features are used to form a fixed-length vector for each modality. Two encoding approaches are adopted respectively: bag-of-words (Sivic and Zisserman, 2003) and MultiVLAD (Jegou and Chum, 2012). The bag-of-words (BoW) representation quantizes local features into visual words using k-means clustering and has been used widely for visual recognition and search (Philbin et al, 2007). The MultiVLAD representation is a variant of the Fisher vector (Perronnin et al, 2010). Two VLAD descriptors obtained from two different codebooks are concatenated, and power-law normalization and PCA are applied to the vector as in (Perronnin et al, 2010). These two encoding approaches differ in that one produces discrete model and the other real-valued. Consequently, we use the Replicated Softmax model for modeling the discrete BoW vectors and the Gaussian RBM for the real-valued MultiVLAD vectors.

3 Models and Algorithms

3.1 Problem description

We aim to learn both a mid-level video representation and a classifier by extending the standard RBM model described in Section 2.1. Given a video dataset $\mathcal{D} = \{(\mathbf{v}_m, y_m)\}_{m=1}^{M}$ with class labels $y \in \{1, ..., C\}$, each video is represented as a N-dimensional low-level feature vector $\mathbf{v}$. Consider modeling videos using a Restricted Boltzmann Machine and let $\mathbf{t}$ denote a F-dimensional latent feature representation of one video. Through training this model, we can map the low-level feature vector $\mathbf{v}$ to the vector $\mathbf{t}$ which can be seen as a mid-level video representation. The learned representation is expected to bridge the semantic gap and improve the effectiveness and efficiency of classification and retrieval. Next we formulate our model with the discrete BoW vectors as model input and omit the case where the real-valued MultiVLAD vectors are used because the formulation is similar.

3.2 Relevance Restricted Boltzmann Machine

The Relevance Restricted Boltzmann Machine (ReRBM) is formulated by integrating sparse Bayesian learning into a

Restricted Boltzmann Machine. The main idea is to jointly learn discriminative latent features as mid-level video representations and a sparse discriminant function as a video classifier.

Let $\mathbf{t}^r = [t_1^r, ..., t_F^r]$ denote a F-dimensional mid-level relevance feature vector of one video. According to Equation 2, the marginal distribution over the BoW vector $\mathbf{v}$ is given by:

$$P(\mathbf{v};\theta) = \frac{1}{\mathcal{Z}(\theta)} \sum_{\mathbf{t}^r} \exp(-E(\mathbf{v}, \mathbf{t}^r; \theta)). \tag{5}$$

Since videos contain more complex and diverse contents than documents, especially when featuring unstructured group activities and events, binary mid-level features learned by the hidden units in the standard RBM are not sufficient to represent the video content. We thus replace the binary hidden units with rectified linear units which are given by:

$$t_j^r = \max(0, t_j),$$
$$P(t_j|\mathbf{v};\theta) = \mathcal{N}(t_j|Kb_j + \sum_{i=1}^{N} W_{ij}v_i, 1), \tag{6}$$

where $\mathcal{N}(\cdot|\mu,\tau)$ denotes a Gaussian distribution with mean μ and variance τ. The rectified linear units taking nonnegative real values can preserve information about relative importance of features. Meanwhile, the rectified Gaussian distribution is semi-conjugate to the Gaussian likelihood. This facilitates the development of variational algorithms for posterior inference and parameter estimation, which will be detailed in Section 3.4.

Let $\boldsymbol{\eta} = \{\boldsymbol{\eta}_y\}_{y=1}^C$ denote a set of class-specific weight vectors. We define the discriminant function as a linear combination of features: $F(y, \mathbf{t}^r, \boldsymbol{\eta}) = \boldsymbol{\eta}_y^T \mathbf{t}^r$. The conditional distribution of classes is defined as follows:

$$P(y|\mathbf{t}^r, \boldsymbol{\eta}) = \frac{\exp(F(y, \mathbf{t}^r, \boldsymbol{\eta}))}{\sum_{y'=1}^{C} \exp(F(y', \mathbf{t}^r, \boldsymbol{\eta}))}, \tag{7}$$

and the classifier is given by:

$$\hat{y} = \arg\max_{y \in C} \mathbb{E}[F(y, \mathbf{t}^r, \boldsymbol{\eta})|\mathbf{v}]. \tag{8}$$

The weights $\boldsymbol{\eta}$ are given a zero-mean Gaussian prior:

$$P(\boldsymbol{\eta}|\boldsymbol{\alpha}) = \prod_{y=1}^{C}\prod_{j=1}^{F} P(\eta_{yj}|\alpha_{yj}) = \prod_{y=1}^{C}\prod_{j=1}^{F} N(\eta_{yj}|0, \alpha_{yj}^{-1}), \tag{9}$$

where $\boldsymbol{\alpha} = \{\boldsymbol{\alpha}_y\}_{y=1}^C$ is a set of hyperparameter vectors, and each hyperparameter α_{yj} is assigned independently to each weight η_{yj}. The hyperpriors over $\boldsymbol{\alpha}$ are given by Gamma distributions:

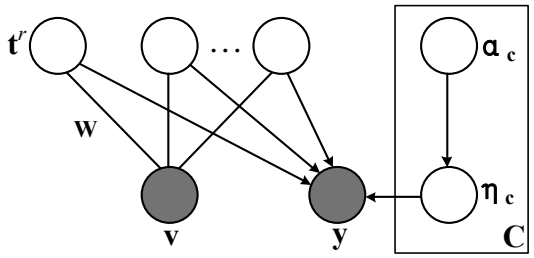


Fig. 3 Relevance Restricted Boltzmann Machine: a hybrid graphical model. The undirected part models the marginal distribution of low-level feature vectors $\mathbf{v}$ and the directed part models the conditional distribution of class labels y given latent features $\mathbf{t}^r$ by using a hierarchical prior on weights $\boldsymbol{\eta}$.

$$P(\boldsymbol{\alpha}) = \prod_{y=1}^{C}\prod_{j=1}^{F} P(\alpha_{yj}) = \prod_{y=1}^{C}\prod_{j=1}^{F} \Gamma(c)^{-1} d^c \alpha_{yj}^{c-1} e^{-d\alpha}, \tag{10}$$

where $\Gamma(c)$ is the Gamma function. To obtain broad hyperpriors, we set c and d to small values, e.g., $c = d = 10^{-4}$. This hierarchical prior, which is a type of automatic relevance determination prior (Neal, 1995), enables the posterior probability of the weights $\boldsymbol{\eta}$ to be concentrated at zero and thus effectively to switch off the corresponding latent features that are considered to be irrelevant to classification. And we refer to those features corresponding to non-zero weights as "class-relevant" features which are discriminative with respect to different video classes.

Finally, given the parameters θ, ReRBM defines the joint distribution:

$$P(\mathbf{v}, y, \mathbf{t}^r, \boldsymbol{\eta}, \boldsymbol{\alpha}; \theta) = P(\mathbf{v};\theta)P(y|\mathbf{t}^r, \boldsymbol{\eta})\left(\prod_{j=1}^{F} P(t_j|\mathbf{v};\theta)\right) \times \left(\prod_{y=1}^{C}\prod_{j=1}^{F} P(\eta_{yj}|\alpha_{yj})P(\alpha_{yj})\right). \tag{11}$$

Figure 3 illustrates ReRBM as a hybrid graphical model with undirected and directed edges. The undirected part models the marginal distribution of video data and the directed part models the conditional distribution of classes given latent features.

3.3 Learning and Inference

To learn a ReRBM, we wish to find parameters $\theta = \{\mathbf{W}, \mathbf{a}, \mathbf{b}\}$ that maximize the log likelihood on $\mathcal{D}$:

$$\log P(\mathcal{D};\theta) = \log \int P(\{\mathbf{v}_m, y_m, \mathbf{t}_m^r\}_{m=1}^M, \eta, \alpha; \theta) d\{\mathbf{t}_m\}_{m=1}^M d\eta d\alpha, \tag{12}$$

and learn the posterior distribution:

$$P(\eta, \alpha|\mathcal{D};\theta) = P(\eta, \alpha, \mathcal{D};\theta)/P(\mathcal{D};\theta). \tag{13}$$

3.3.1 Variational bounds

Since exactly computing $P(\mathcal{D};\theta)$ is intractable, we employ variational methods to optimize a lower bound $\mathcal{L}$ on the log likelihood by introducing a variational distribution to approximate $P(\{\mathbf{t}_m\}_{m=1}^M, \eta, \alpha|\mathcal{D};\theta)$:

$$Q(\{\mathbf{t}_m\}_{m=1}^M, \eta, \alpha) = \left(\prod_{m=1}^M \prod_{j=1}^F q(t_{mj})\right) q(\eta) q(\alpha). \tag{14}$$

Using the Jensens inequality, we have:

$$\log P(\mathcal{D};\theta) \geqslant \int Q(\{\mathbf{t}_m\}_{m=1}^M, \eta, \alpha) \times \log \frac{P(\{\mathbf{v}_m, y_m, \mathbf{t}_m^r\}_{m=1}^M, \eta, \alpha; \theta)}{Q(\{\mathbf{t}_m\}_{m=1}^M, \eta, \alpha)} d\{\mathbf{t}_m\}_{m=1}^M d\eta d\alpha, \tag{15}$$

where

$$P(\{\mathbf{v}_m, y_m, \mathbf{t}_m^r\}_{m=1}^M, \eta, \alpha; \theta) = \left(\prod_{m=1}^M P(\mathbf{v}_m;\theta) P(y_m|\mathbf{t}_m^r, \eta) P(\mathbf{t}_m|\mathbf{v}_m;\theta)\right) P(\eta|\alpha) P(\alpha).$$

Note that $P(y_m|\mathbf{t}_m^r, \eta)$ is not conjugate to the Gaussian prior, which makes it intractable to compute the variational factors $q(\eta)$ and $q(t_{mj})$. Here we use a quadratic bound on the log-sum-exp (LSE) function (Bohning, 1992) to derive a further bound. We rewrite $P(y_m|\mathbf{t}_m^r, \eta)$ as follows:

$$P(y_m|\mathbf{t}_m^r, \eta) = \exp(\mathbf{y}_m^T \mathbf{T}_m^r \eta - \text{lse}(\mathbf{T}_m^r \eta)), \tag{16}$$

where $\mathbf{T}_m^r \eta = [(\mathbf{t}_m^r)^T \eta_1, ..., (\mathbf{t}_m^r)^T \eta_{C-1}]$, $\mathbf{y}_m = \mathbb{I}(y_m = c)$ is the one-of-C encoding of class label y_m and $\text{lse}(\mathbf{x}) \triangleq \log(1 + \sum_{y'=1}^{C-1} \exp(x_{y'}))$ (we set $\eta_C = \mathbf{0}$ to ensure identifiability). In (Bohning, 1992), the LSE function is expanded as a second order Taylor series around a point φ, and an upper bound is found by replacing the Hessian matrix $\mathbf{H}(\varphi)$ with a fixed matrix $\mathbf{A} = \frac{1}{2}[\mathbf{I}_{C^*} - \frac{1}{C^*+1}\mathbf{1}_{C^*}\mathbf{1}_{C^*}^T]$ such that $\mathbf{A} \succ \mathbf{H}(\varphi)$, where $C^* = C-1$, $\mathbf{I}_{C^*}$ is the identity matrix of size $M \times M$ and $\mathbf{1}_{C^*}$ is a M-vector of ones. Thus, similar to (Murphy, 2012), we have:

$$\log P(y_m|\mathbf{t}_m^r, \eta) \geqslant J(y_m, \mathbf{t}_m^r, \eta, \varphi_m) = \mathbf{y}_m^T \mathbf{T}_m^r \eta - \frac{1}{2}(\mathbf{T}_m^r \eta)^T \mathbf{A} \mathbf{T}_m^r \eta + \mathbf{s}_m^T \mathbf{T}_m^r \eta - \kappa_i, \tag{17}$$

$$\mathbf{s}_m = \mathbf{A}\varphi_m - \exp(\varphi_m - \text{lse}(\varphi_m)), \tag{18}$$

$$\kappa_i = \frac{1}{2}\varphi_m^T \mathbf{A} \varphi_m - \varphi_m^T \exp(\varphi_m - \text{lse}(\varphi_m)) + \text{lse}(\varphi_m), \tag{19}$$

where $\varphi_m \in \mathbb{R}^{C^*}$ is a vector of variational parameters. Substituting $J(y_m, \mathbf{t}_m^r, \eta, \varphi_m)$ into Equation 11, we can obtain a further lower bound:

$$\log P(\mathcal{D};\theta) \geqslant \mathcal{L}(\theta, \varphi) = \sum_{m=1}^M \log P(\mathbf{v}_m;\theta) + \mathbb{E}_Q\Big[\sum_{m=1}^M J(y_m, \mathbf{t}_m^r, \eta, \varphi_m) + \sum_{m=1}^M \log P(\mathbf{t}_m|\mathbf{v}_m;\theta) + \log P(\eta|\alpha) + \log P(\alpha) - Q(\{\mathbf{t}_m\}_{m=1}^M, \eta, \alpha)\Big]. \tag{20}$$

Now we have converted the problem of model learning into maximizing the lower bound $\mathcal{L}(\theta, \varphi)$ with respect to the variational posteriors $q(\eta)$, $q(\alpha)$ and $q(\mathbf{t}) = \{q(t_{mj})\}$ as well as the parameters θ and $\varphi = \{\varphi_m\}$. We can give some insights into the objective function $\mathcal{L}(\theta, \varphi)$: the first term is exactly the marginal log likelihood of video data and the second term is a variational bound of the conditional log likelihood of classes. Thus maximizing $\mathcal{L}(\theta, \varphi)$ is equivalent to finding a set of model parameters and latent features which could fit video data well and simultaneously make good predictions for the class each video belongs to.

3.3.2 Variational inference

Due to the conjugacy properties of the chosen distributions, we can directly calculate free-form variational posteriors $q(\eta)$, $q(\alpha)$ and parameters φ:

$$q(\eta) = \mathcal{N}(\eta|\mathbf{E}_\eta, \mathbf{V}_\eta), \tag{21}$$

$$q(\alpha) = \prod_{y=1}^C \prod_{j=1}^F \text{Gamma}(\alpha_{yj}|\hat{c}, \hat{d}_{yj}), \tag{22}$$

$$\varphi_m = \langle \mathbf{T}_m^r \rangle_{q(\mathbf{t})} \mathbf{E}_\eta, \tag{23}$$

where $\langle\cdot\rangle_q$ denotes an exception with respect to the distribution q and

$$\mathbf{V}_\eta = \left(\sum_{m=1}^{M} \left\langle (\mathbf{T}_m^r)^T \mathbf{A} \mathbf{T}_m^r \right\rangle_{q(\mathbf{t})} + \text{diag}\langle \alpha_{yj} \rangle_{q(\alpha)}\right)^{-1}, \tag{24}$$

$$\mathbf{E}_\eta = \mathbf{V}_\eta \sum_{m=1}^{M} \left\langle (\mathbf{T}_m^r)^T \right\rangle_{q(\mathbf{t})} (\mathbf{y}_m + \mathbf{s}_m), \tag{25}$$

$$\hat{c} = c + \frac{1}{2}, \tag{26}$$

$$\hat{d}_{yj} = d + \frac{1}{2}\langle \eta_{yj}^2 \rangle_{q(\eta)}. \tag{27}$$

For $q(\mathbf{t})$, the calculation is not directly implemented because of the rectification. Inspired by (Harva and Kaban, 2007), we have the following free-form solution:

$$q(t_{mj}) = \frac{\omega_{pos}}{Z}\mathcal{N}(t_{mj}|\mu_{pos}, \sigma_{pos}^2)u(t_{mj}) + \frac{\omega_{neg}}{Z}\mathcal{N}(t_{mj}|\mu_{neg}, \sigma_{neg}^2)u(-t_{mj}), \tag{28}$$

where $u(\cdot)$ is the unit step function. See Appendix A for a detailed descripton on how the parameters of $q(t_{mj})$ are estimated.

Given θ, through repeating the updates of Equations 18-20 and 23 to maximize $\mathcal{L}(\theta, \varphi)$, we can obtain the variational posteriors $q(\eta)$, $q(\alpha)$ and $q(\mathbf{t})$.

3.3.3 Parameter estimation

After the variational posteriors $q(\eta)$, $q(\alpha)$ and $q(\mathbf{t})$ are computed, we estimate model parameters $\theta = \{\mathbf{W}, \mathbf{a}, \mathbf{b}\}$ by using stochastic gradient descent to maximize $\mathcal{L}(\theta, \varphi)$. The derivatives of $\mathcal{L}(\theta, \varphi)$ with respect to θ are given by:

$$\frac{\partial \mathcal{L}(\theta, \varphi)}{\partial W_{ij}} = \left\langle v_i t_j^r \right\rangle_{data} - \left\langle v_i t_j^r \right\rangle_{model} + \frac{1}{M}\sum_{m=1}^{M} v_{mi}\left(\langle t_{mj}\rangle_{q(\mathbf{t})} - \sum_{i=1}^{N} W_{ij} v_{mi} - K b_j\right), \tag{29}$$

$$\frac{\partial \mathcal{L}(\theta, \varphi)}{\partial a_i} = \langle v_i \rangle_{data} - \langle v_i \rangle_{model}, \tag{30}$$

$$\frac{\partial \mathcal{L}(\theta, \varphi)}{\partial b_j} = \left\langle t_j^r \right\rangle_{data} - \left\langle t_j^r \right\rangle_{model} + \frac{K}{M}\sum_{m=1}^{M}\left(\langle t_{mj}\rangle_{q(\mathbf{t})} - \sum_{i=1}^{N} W_{ij} v_{mi} - K b_j\right), \tag{31}$$

where the derivatives of $\sum_{m=1}^{M} \log P(\mathbf{v}_m; \theta)$ are the same as those in (Salakhutdinov and Hinton, 2009).

This leads to the following variational EM algorithm. E-step: Calculate variational posteriors $q(\eta)$, $q(\alpha)$ and $q(\mathbf{t})$. M-step: Estimate parameters $\theta = \{\mathbf{W}, \mathbf{a}, \mathbf{b}\}$ through maximizing $\mathcal{L}(\theta, \varphi)$. These two steps are repeated until $\mathcal{L}(\theta, \varphi)$ converges. The entire learning procedure is summarized in Algorithm 1. After the learning is completed, the prediction for new videos can be easily obtained via Equation 8:

$$\hat{y} = \arg\max_{y \in C} \left\langle \eta_y^T \right\rangle_{q(\eta)} \langle \mathbf{t}^r \rangle_{p(\mathbf{t}|\mathbf{v};\theta)}. \tag{32}$$

According to Equation 6, the computational complexity for the inference of latent features $\mathbf{t}$ is $\mathcal{O}(NF)$. Thus the computation for the prediction is $\mathcal{O}(NF^2)$ (typically $F \ll N$) which is linearly proportional to the dimensionality of the model input.

Algorithm 1 Variational EM for learning ReRBM

Input:
Video dataset $\mathcal{D} = \{(\mathbf{v}_m, y_m)\}_{m=1}^M$
Output:
Model parameters $\theta = \{\mathbf{W}, \mathbf{a}, \mathbf{b}\}$
Variational posteriors $q(\eta)$, $q(\alpha)$ and $q(\mathbf{t})$
Initialize θ
repeat
  for $m = 1$ **to** M **do**
    Update $q(\mathbf{t}_m)$ as in Equation 28
    Update φ_m as in Equation 23
  end for
  Update $q(\alpha)$ as in Equation 22
  Update $q(\eta)$ as in Equation 21
  Optimize $\mathbf{W}$, $\mathbf{a}$ and $\mathbf{b}$ with stochastic gradient descent using Equation 29-31
until convergence

3.4 Extension to multiple modalities

The formulation so far is limited to a single BoW vector from a single modality. Single-modality features are usually limited for videos containing complicated content. In particular, both visual and audio features are often necessary. For example, the tune of the happy birthday song is a very useful feature for distinguishing a birthday party from a normal dinner party. Therefore in this work, in addition to motion features typically used for video representation, static appearance and auditory features are also used as described in Section 2.2. More specifically, to sufficiently characterise complex videos, we consider using features of three modalities, i.e., SIFT, STIP and MFCC. Accordingly we extend ReRBM to Multimodal ReRBM to discover a unified mid-level feature representation from multi-modal low-level feature inputs. As shown in Figure 4, our Multimodal ReRBM uses the undirected part to model the multimodal data $\mathbf{v} = \{\mathbf{v}^{\text{modl}}\}_{l=1}^L$. Consequently, its joint distribu-

tion can be given by replacing $P(\mathbf{v};\theta)$ in Equation 11 with $\prod_{l=1}^{L} P(\mathbf{v}^{\text{mod}l};\theta^{\text{mod}l})$.

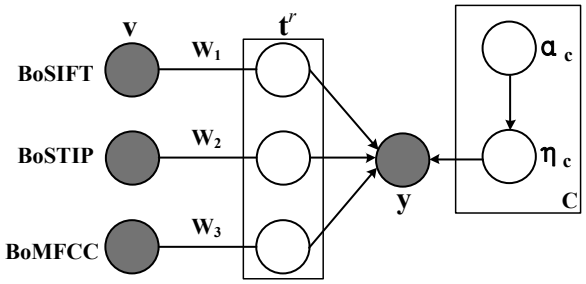


Fig. 4 Multimodal ReRBM: the extension of ReRBM for modeling multi-modal input data including bag-of-SIFT, bag-of-STIP and bag-of-MFCC.

For learning Multimodal ReRBM, we just need to additionally calculate the gradients of $\theta^{\text{mod}l}$ for each modality l and estimate $\theta^{\text{mod}l}$ with stochastic gradient descent in the M-step while other updating rules remain unchanged.

4 Experiments

4.1 Datasets and settings

For evaluating the performance of the learned mid-level video representation, we test our models on the Unstructured Social Activity Attribute (USAA) dataset[3] for group activity recognition and the Event Video (EVVE) dataset[4] for event retrieval. We also present quantitative and qualitative comparisons with other supervised latent variable models (namely MedLDA and sRBM) and some other baselines when appropriate on these datasets. In all experiments, the contrastive divergence is used to efficiently approximate the derivatives of the marginal log likelihood and the unsupervised training on RBM is used to initialize θ.

The USAA dataset consists of 8 semantic classes of social activity videos collected from the Internet. The eight classes are: birthday party, graduation party, music performance, non-music performance, parade, wedding ceremony, wedding dance and wedding reception. The dataset contains a total of 1466 videos and approximate 100 videos per-class for training and testing respectively. These videos last from 20 seconds to 8 minutes with an average duration of 3 minutes and contain very complex and diverse contents, which brings significant challenges for content analysis. Some example video frames are shown in Figure 5. For each video, three local features (SIFT, STIP and MFCC) are extracted and result in three bag-of-words vectors (5000 dimensions for SIFT and STIP, and 4000 dimensions for MFCC) using a soft-weighting clustering algorithm.

The EVVE dataset contains 2375 database videos (1123 negatives which do not belong to any events) and 620 query videos which were returned to 13 different queries on YouTube. The total length of the videos is 166 hours. Different from event detection tasks which aim to recognize video event categories, it is dedicated to the retrieval of particular events, as illustrated in Figure 6. The query phrases include "Austerity riots in Barcelona in 2012", "Concert of Die toten Hosen in 2012" and "Egyptian revolution: Tahrir Square demonstrations", etc. EVVE also includes some negative samples which are relevant events but they took place not in the same place or time, such as riots occurring in different places but not in Barcelona 2012. All videos are sampled at a fixed rate of 15 fps and resized to a maximum of 120k pixels. Square-root SIFT features are extracted for each frame on a dense grid. Then the SIFT features of a frame are encoded into a MultiVLAD vector. For characterizing the entire video, the Mean-MultiVLAD (MMV) representation is obtained by averaging all the frame descriptors.

4.2 Experiments on group activity recognition

4.2.1 Comparisons against alternative latent variable models

To verify the discriminative power of the class-relevant features learned by our ReRBM, we present quantitative classification results compared with other supervised latent variable models namely MedLDA (Zhu et al, 2012) and sRBM (Larochelle et al, 2012) in the case of different modalities on the USAA dataset. We have tried our best to tune these compared models and report the best results.

Figure 7 shows the classification accuracy of different models for three single-modal local features: SIFT, STIP and MFCC. We test different latent feature dimensions (corresponding to the number of hidden units/variables in different models) from 20 to 60. We can see that ReRBM achieves higher classification accuracy than MedLDA and sRBM in all cases regardless what low-level feature modality or latent feature dimension is used. This result demonstrates that ReRBM can find more discriminative representations for complex video data through leveraging sparse Bayesian learning to incorporate class label information into representation learning. The sparsity of classifier weights effectively selects latent features that are relevant to the class labels. The performance of ReRBM in the case of high dimensional mid-level feature representation is slightly worse than the case of the low dimension because no more class-relevant features are learned when the dimension of the latent features continues to increase. Note that the features learned from SIFT perform better than those learned from

[3] http://www.eecs.qmul.ac.uk/~yf300/USAA/download
[4] http://pascal.inrialpes.fr/data/evve

(a) Wedding Dance (b) Birthday Party (c) Graduation Ceremony

Fig. 5 Example videos of the "Wedding Dance", "Birthday Party" and "Graduation Ceremony" classes from the USAA dataset.

(a) Austerity riots in Barcelona in 2012

(b) Concert of Die toten Hosen in 2012

(c) Egyptian revolution: Tahrir Square demonstrations

Fig. 6 Example videos of the events "Austerity riots in Barcelona in 2012", "Concert of Die toten Hosen in 2012" and "Egyptian revolution: Tahrir Square demonstrations" from the EVVE dataset.

the other two local features which suggests that scene and object information are more useful for understanding complex videos.

4.2.2 Comparisons against other baselines

We compare Multimodal ReRBM with the baselines in (Fu et al, 2012) which reports the state of the art results on the USAA dataset. More specifically, we compare ReRBM with three models in (Fu et al, 2012) given 10 and 100 instances per class respectively for model training. They are:

- **Direct:** Direct SVM or KNN classification on raw video BoW vectors (14000 dimensions obtained by concatenating the SIFT, STIP and MFCC BoW vectors). SVM is used for experiments with 100 instances per class and KNN with 10 instances.
- **SVM-UD+LR:** SVM attribute classifiers are first trained for the 69 human-defined attributes, and then a logistic regression (LR) classifier is trained using the attribute classifier outputs as mid-level video representation. Note that additional annotations in the form of human-defined instance-level attribute vectors are used in this model, giving it a unfair advantage over our ReRBM.
- **SLAS+LR:** Semi-latent attribute space is learned, and then a LR classifier is trained using the aggregation of 69 human-defined, 8 class-conditional and 8 latent topics as mid-level video representation. Again additional manual annotations of instance-level attributes are used.

In addition, our ReRBM is compared with another baseline not reported in (Fu et al, 2012). In this baseline, multimodal features extracted by Replicated Softmax, another undirected topic model, are connected together as video representations, followed by learning a multi-class SVM classifier from the representations (Tsochantaridis et al, 2004). This baseline is denoted as **RS+SVM**. Similar to SVM-UD+LR and SLAS+LR, this is a two-staged approach with the tasks of mid-level representation learning and classifier learning tackled separately, in contrast to our unified model.

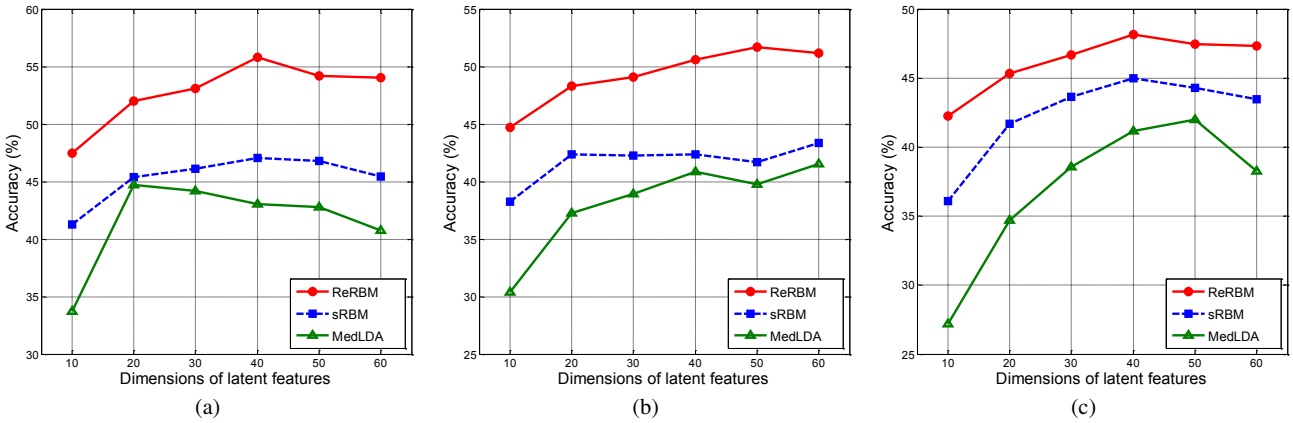


Fig. 7 Classification accuracy of different supervised latent variable models for single-modal local features (a) SIFT, (b) STIP and (c) MFCC.

Table 1 Classification accuracy of different methods for multimodal features.

Method		Multimodal ReRBM		RS+SVM	Direct	SVM-UD+LR	SLAS+LR
Accuracy (%)	100 Inst	60-D	60.22	54.60	**66.0** (14000-D)	65.0 (69-D)	65.0 (85-D)
		90-D	62.69	56.10			
		120-D	63.79	57.34			
		150-D	64.06	59.26			
		180-D	64.72	60.63			
	10 Inst	60-D	38.68	23.73	29.0 (14000-D)	37.0 (69-D)	40.0 (85-D)
		90-D	**41.29**	28.53			
		120-D	**43.48**	30.59			
		150-D	**43.72**	33.47			
		180-D	**44.99**	35.94			

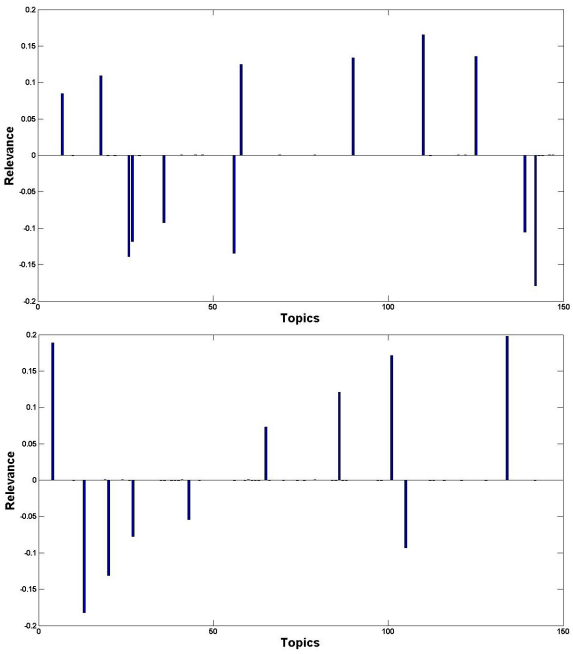
The results are shown in Table 2. Here the dimensionality of latent features for each modality is assumed to be the same, ranging from 20 to 60 (totally from 60 to 180). When the labeled training data is plentiful (100 instances per class), Table 2 shows that the classification performance of our Multimodal ReRBM is similar to the three baselines in (Fu et al, 2012) but clearly better than RS+SVM. Note that both SVM-UD+LR and SLAS+LR use human defined concepts which need additional label information and thus has an unfair advantage. In contrast, our model does not require an attribute ontology and manual annotation of attribute vectors, and is thus able to be better generalized to large scale problem. Since the Direct method (with original dimensions of low-level feature representations) performs strongly, this result suggests that learning mid-level representations are not necessarily useful given sufficient training data. However, the Direct method is with high dimension, and will cost huge computation in practical applications.

When considering the classification scenario where only a very small number of training data are available (10 instances per class), Multimodal ReRBM outperforms all four baselines with the number of latent features more than 90. Again this result demonstrates the importance of learning sparse and discriminative mid-level features. In particular, the sparsity of the mid-level features learned by ReRBM can effectively prevent overfitting to specific training instances given limited training data. It is also noted that our model outperforms RS+SVM in both cases, which demonstrates the advantage of jointly learning latent features and classifier weights through sparse Bayesian learning over a two-staged approach.

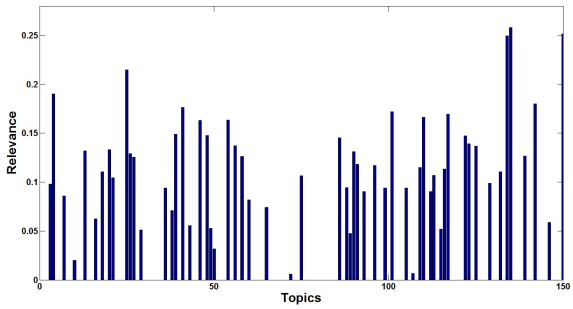
4.2.3 Further evaluation on learned features

To validate the sparsity of the class-relevant features learned by ReRBM, Figure 8(a) illustrates the degree of relevance between features and two different classes. We can see that the learned class-relevant features are very sparse and distinct between two classes; they thus are able to provide discriminative information for distinguishing these two classes. We also show the average relevance of features on all 8 classes in Figure 8(b). The overall sparsity can be also observed, which leads to good generalization for new instances and robustness given small training datasets.

To gain some insight into what the learned class-relevant features actually correspond to in videos, we visualize some of them for static appearance (SIFT) and motion (STIP) in



(a) Class-relevant features for two different classes.



(b) Class-relevant features averaged on all classes.

Fig. 8 Class-relevant features discovered by ReRBM. Vertical axis indicates the degree of relevance.

Figure 9(a) and (b) respectively. We also visualize latent features learned by sRBM in Figure 9(c) for comparison. It can be seen that the local low-level features included in the class-relevant features mostly lie on people and objects which are relevant to video classes (e.g., for SIFT the cake in the *birthday party* video, and the bride and flowers in the *wedding dance* video; for STIP the performer in the *music performance* video and troops in the *parade* video). Some negative class-relevant features are also shown, e.g., the speaker's desk which usually dose not appear in a birthday party, and blowing candles which seldom happens in a music performance. In the meantime, those from the non-relevant features mostly lie on backgrounds which are shared between classes (e.g., trees, windows and floors). This qualitative result suggests that although there are no human annotations available, ReRBM can still automatically discover semantic and discriminative visual patterns in videos. We also observe that ReRBM learns more discriminative features than sRBM because only very sparse class-relevant features are permitted to contribute to the decision of classification in ReRBM, which enables the discriminative semantic information to be concentrated on a subset of the learned latent features.

4.2.4 Scalability of the inference

Although the inference of latent features in the learning procedure is difficult due to the explaining away effect caused by the direct part, in the prediction stage because the class labels are unknown, the latent features are conditional independent given the low-level features and can be directly calculated by Equation 6 efficiently as described in Section 3.3.3. This is what we want since the learning of the model can be completed offline but the prediction is usually online and need be real-time. We evaluate the inference time of Multimodal ReRBM on the test set of the USAA dataset. The time of computing 180-D latent features is 0.14s using Matlab on a standard desktop with a 3.10GHz Intel Core processor.

4.3 Experiments on event retrieval

To further validate the semantic knowledge discovered by the class-relevant features, we evaluate the retrieval performance of ReRBM on the EVVE dataset and compare it with sRBM, as well as the MMV representation which is the low-level feature representation provided by (Revaud et al, 2013). Note that the MMV features, which are also used as model input for both ReRBM and sRBM, have a dimensionality of 1024 which is significantly higher than the dimensionality of the learned mid-level features by ReRBM and sRBM. For ReRBM, we also present the retrieval results of using all learned features and only relevance features respectively to show the effectiveness of sparse Bayesian learning. We use the database videos (only positive samples) as the training set and the query videos as the test sets. Similar to the setting in (Salakhutdinov and Hinton, 2009), a retrieval video is considered relevant to the query video if they have the same class label. The similarity is measured using the Cosine distance between the feature vectors representing the video content. To measure the retrieval performance, the mean Average Precision (mAP) is computed per event given different mid-level feature dimensions for ReRBM and sRBM. To evaluate the overall performance, the average of the mAPs over the 13 different events (avg-mAP) is also given.

The results in Table 3 show clearly that the mid-level feature representation learned by our ReRBM achieves the highest average mAP than using the low-level MMV feature directly. This indicates that the mid-level features learned by ReRBM can effectively bridge the semantic gap. Interestingly, using only the class-relevant features performs bet-

(a) Latent features learned by ReRBM for static appearance (SIFT). Left: Birthday party; Right: Wedding dance

(b) Latent features learned by ReRBM for motion (STIP). Left: Music performance; Right: Parade

(c) Latent features learned by sRBM for SIFT (Left) and STIP (Right)

Fig. 9 Visualization of learned latent features. Red (blue) circles indicate local low-level features included in positive (negative) class-relevant features learned by ReRBM and green crosses indicate those included in irrelevant features learned by ReRBM; red crosses indicate local low-level features included in relevant features learned by sRBM.

ter than using all latent features learned by ReRBM. This suggests that the class-relevant features do capture semantic information related to classes thus more useful for retrieval, whilst the non-relevant features are non-discriminative thus are distracted to include. Furthermore, with the growth of the dimensionality, more non-relevance features will be learned than relevance features due to the sparsity constraint enforced on the learned features in ReRBM. Thus in higher dimensions using all learned features has worse performance. In contrast, when only relevant features are used the perform keeps increasing.

Figure 10 depicts the quality of retrieval results for two example queries, where sample frames of the top-5 retrieved videos are shown. It can be observed that ReRBM retrieves videos which belong to the same target event very well; sRBM can find some semantically related videos (e.g., some riot events but not in Barcelona), while MMV only captures visual similarity between videos which leads to poor retrieval performance as many different classes of events share visual similarity (e.g. many different events have large crowds).

5 Conclusion

We have proposed a supervised Restricted Boltzmann Machine, called Relevance Restricted Boltzmann Machine (ReRBM), to learn discriminative mid-level feature representations for the classification and retrieval of unstructured group activities and events in videos. In ReRBM, sparse Bayisian learning is integrated with the RBM to discover sparse class-relevant features. Rectified linear units are employed in the place of binary hidden units to better describe complex video content and enable an efficient variational method to be developed for parameter estimation and inference. ReRBM can be readily be extended to take multimodal data as inputs. Our experimental results demonstrated that ReRBM is able to learn a more discriminative video representation than other supervised latent variable models and achieves state of the art classification and retrieval performance, particularly given insufficient training data.

In this work, we have only considered a single layer RBM. RBM has been constructed with a deep learning structure with multiple layers (Hinton et al, 2006). Thus it seems a promising direction to generalize our framework of representation learning to a multilayer structure. Furthermore, when combined with other deep learning models such

Table 2 Results on retrieval performance (mAP) using different feaurs.

Event number	sRBM 20-D	sRBM 40-D	sRBM 60-D	ReRBM 20-D	ReRBM 40-D	ReRBM 60-D	ReRBM + Relevance Features 20-D	ReRBM + Relevance Features 40-D	ReRBM + Relevance Features 60-D	MMV 1024-D
#1	0.810	0.718	0.715	0.882	0.842	0.747	**0.891**	0.862	0.867	0.555
#2	0.538	0.522	0.559	0.673	0.712	0.691	0.672	0.740	**0.749**	0.344
#3	0.203	0.127	0.124	0.215	0.354	0.319	0.212	**0.405**	0.352	0.092
#4	0.473	0.532	0.535	0.777	0.753	0.734	0.788	0.780	**0.810**	0.467
#5	0.317	0.310	0.310	0.477	0.448	0.444	0.478	0.470	**0.528**	0.238
#6	0.215	0.296	0.326	0.477	0.469	0.458	0.476	0.519	**0.659**	0.264
#7	0.219	0.240	0.265	0.410	0.377	0.377	0.405	0.444	**0.539**	0.208
#8	0.146	0.175	0.184	0.292	0.254	0.258	0.296	0.261	**0.347**	0.120
#9	0.228	0.172	0.165	0.540	0.507	0.504	0.551	0.525	**0.581**	0.123
#10	0.553	0.571	0.569	0.771	0.664	0.808	0.776	0.601	**0.869**	0.365
#11	0.306	0.331	0.326	0.503	0.538	0.519	0.530	0.597	**0.699**	0.257
#12	0.880	0.848	0.820	0.905	0.888	0.899	**0.906**	0.856	0.879	0.759
#13	0.619	0.842	0.857	**0.926**	0.869	0.801	0.923	0.874	0.837	0.601
avg-mAP	0.424	0.437	0.443	0.604	0.590	0.581	0.608	0.610	**0.671**	0.338

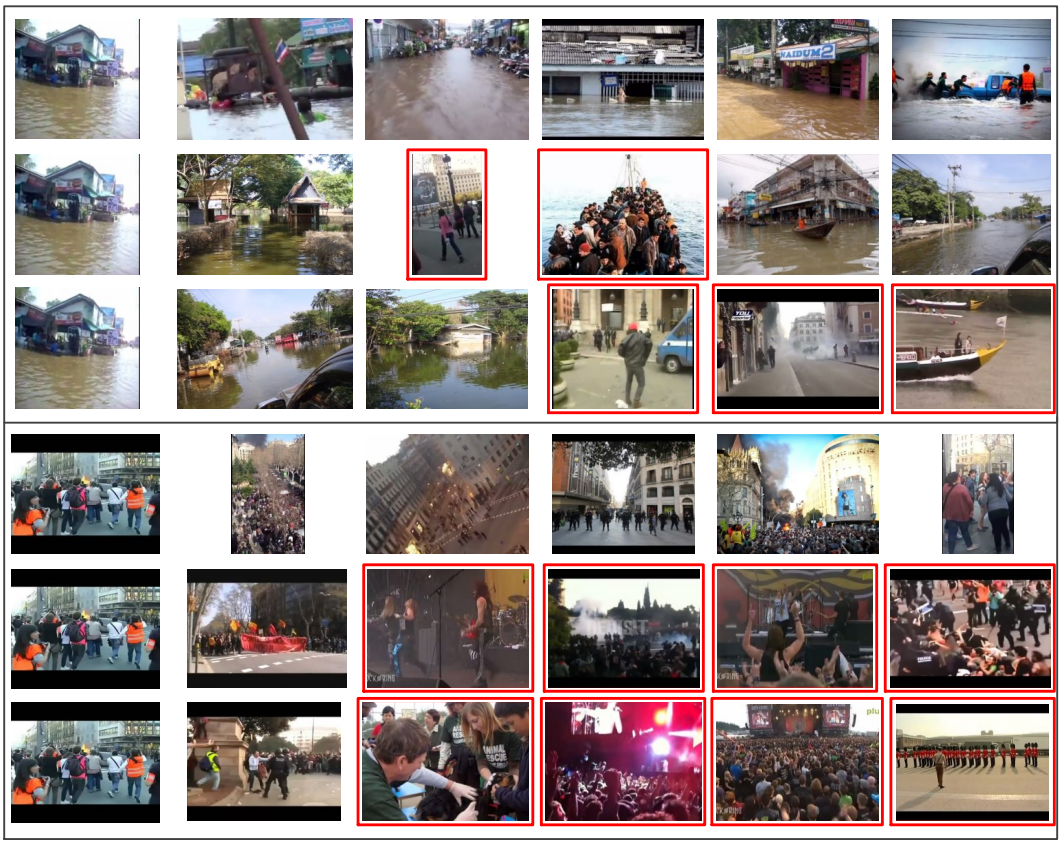


Fig. 10 Sample retrieval results using ReRBM + Relevance Features (first row), sRBM (second row) and MMV (third row). The left item in the every row is the query video, and red rectangles indicate mistakes (best viewed in color).

as deep convolutional neural networks (Hinton et al, 2006; Sun et al, 2013), it is possible to learn from the low-level feature representation at the bottom all the way up to the class labels at the top in a single unified model. This is another interesting direction and part of the on-going work.

Appendix A. Parameters of free-form variational posterior $q(t_{mj})$

The expressions of parameters in $q(t_{mj})$ (Equation 28) are listed as follows:

$$\omega_{pos} = \mathcal{N}(\alpha|\beta, \gamma+1),\ \sigma^2_{pos} = (\gamma^{-1}+1)^{-1}, \tag{33}$$

$$\mu_{pos} = \sigma^2_{pos}(\frac{\alpha}{\gamma}+\beta), \tag{34}$$

$$\omega_{neg} = \mathcal{N}(\alpha|0, \gamma),\ \sigma^2_{neg} = 1,\ \mu_{neg} = \beta, \tag{35}$$

$$Z = \frac{1}{2}\omega_{pos}\text{erfc}\left(\frac{-\mu_{pos}}{\sqrt{2\sigma^2_{pos}}}\right) + \frac{1}{2}\omega_{neg}\text{erfc}\left(\frac{\mu_{neg}}{\sqrt{2\sigma^2_{neg}}}\right), \tag{36}$$

where erfc$(\cdot)$ is the complementary error function and

$$\alpha = \left\langle \frac{\boldsymbol{\eta}_{\cdot j}\left(\mathbf{y}_m + \mathbf{s}_m - \sum_{j'\neq j}\boldsymbol{\eta}_{\cdot j'}\mathbf{A}t^r_{mj'}\right)}{\boldsymbol{\eta}_{\cdot j}\mathbf{A}\boldsymbol{\eta}^T_{\cdot j}} \right\rangle_{q(\boldsymbol{\eta})q(\mathbf{t})}, \quad (37)$$

$$\gamma = \left\langle \boldsymbol{\eta}_{\cdot j}\mathbf{A}\boldsymbol{\eta}^T_{\cdot j} \right\rangle^{-1}_{q(\boldsymbol{\eta})}, \ \beta = \sum_{i=1}^{N} W_{ij}v_{mi} + Kb_j. \quad (38)$$

We can see that $q(t_{mj})$ depends on expectations over $\boldsymbol{\eta}$ and $\{t_{mj'}\}_{j'\neq j}$, which is consistent with the graphical model representation of ReRBM in Figure 3.

References

Bengio Y, Courville AC, Vincent P (2012) Unsupervised feature learning and deep learning: A review and new perspectives. CoRR abs/1206.5538

Blei D, Ng A, Jordan M (2003) Latent dirichlet allocation. The Journal of Machine Learning Research 3:993–1022

Bohning D (1992) Multinomial logistic regression algorithm. Annals of the Institute of Statistical Mathematics 44(1):197–200

Desjardins G, Courville AC, Bengio Y (2012) On training deep boltzmann machines. CoRR abs/1203.4416

Farhadi A, Endres I, Hoiem D, Forsyth D (2009) Describing objects by their attributes. In: Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on, IEEE, pp 1778–1785

Fu Y, Hospedales T, Xiang T, Gong S (2012) Attribute learning for understanding unstructured social activity. In: Computer Vision-CECCV 2012, Springer, pp 530–543

Gopalan R (2013) Joint sparsity-based representation and analysis of unconstrained activities. In: Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on, IEEE, pp 2738–2745

Harva M, Kaban A (2007) Variational learning for rectified factor analysis. Signal Processing 87(3):509–527

Hinton G (2002) Training products of experts by minimizing contrastive divergence. Neural Computation 14(8):1771–1800

Hinton G, Salakhutdinov R (2006) Reducing the dimensionality of data with neural networks. Science 313(5786):504–507

Hinton G, Osindero S, Teh Y (2006) A fast learning algorithm for deep belief nets. Neural Computation 18(7):1527–1554

Izadinia H, Shah M (2012) Recognizing complex events using large margin joint low-level event model. In: Computer VisionCECCV 2012, Springer, pp 430–444

Jegou H, Chum O (2012) Negative evidences and co-occurences in image retrieval: The benefit of pca and whitening. In: Computer VisionCECCV 2012, Springer, pp 774–787

Lampert C, Nickisch H, Harmeling S (2013) Attribute-based classification for zero-shot learning of object categories. Pattern Analysis and Machine Intelligence, IEEE Transactions on 36(3):453 – 465

Lampert CH, Nickisch H, Harmeling S (2009) Learning to detect unseen object classes by between-class attribute transfer. In: Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on, IEEE, pp 951–958

Laptev I (2005) On space-time interest points. International Journal of Computer Vision 64(2-3):107–123

Laptev I, Marszalek M, Schmid C, Rozenfeld B (2008) Learning realistic human actions from movies. In: Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on, IEEE, pp 1–8

Larochelle H, Mandel M, Pascanu R, Bengio Y (2012) Learning algorithms for the classification restricted boltzmann machine. The Journal of Machine Learning Research 13:643–669

Liu J, Kuipers B, Savarese S (2011) Recognizing human actions by attributes. In: Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on, IEEE, pp 3337–3344

Logan B (2000) Mel frequency cepstral coefficients for music modeling. In: International Symposium on Music Information Retrieval

Lowe D (2004) Distinctive image features from scale-invariant keypoints. International Journal of Computer Vision 60(2):91–110

Murphy K (2012) Machine learning: a probabilistic perspective. MIT Press

Nair V, Hinton G (2010) Rectified linear units improve restricted boltzmann machines. In: Proceedings of the 27th International Conference on Machine Learning, ACM, pp 807–814

Neal R (1995) Bayesian learning for neural networks. PhD thesis, University of Toronto

Perronnin F, Sanchez J, Mensink T (2010) Improving the fisher kernel for large-scale image classification. In: Computer VisionCECCV 2010, Springer, pp 143–156

Philbin J, Chum O, Isard M, Sivic J, Zisserman A (2007) Object retrieval with large vocabulary and fast spatial matching. In: Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on, IEEE, pp 1–8

Ranzato M, Hinton GE (2010) Modeling pixel means and covariances using factorized third-order boltzmann machines. In: The Twenty-Third IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2010, San Francisco, CA, USA, 13-18 June 2010, pp 2551–2558

Rasiwasia N, Vasconcelos N (2013) Latent dirichlet allocation models for image classification. Pattern Analysis and Machine Intelligence, IEEE Transactions on 35(11):2665–2679

Reddy K, Shah M (2013) Recognizing 50 human action categories of web videos. Machine Vision and Applications 24(5):971–981

Revaud J, Douze M, Schmid C, Jegou H (2013) Event retrieval in large video collections with circulant temporal encoding. In: Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on, IEEE, pp 2459–2466

Rodriguez M, Ahmed J, Shah M (2008) Action mach: a spatio-temporal maximum average correlation height filter for action recognition. In: Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on, IEEE, pp 1–8

Salakhutdinov R, Hinton G (2009) Replicated softmax: an undirected topic model. In: Advances in Neural Information Processing Systems, MIT Press, vol 22, pp 1607–1614

Schuldt C, Laptev I, Caputo B (2004) Recognizing human actions: a local svm approach. In: Pattern Recognition, 2004. ICPR 2004. Proceedings of the 17th International Conference on, IEEE, vol 3, pp 32–36

Sivic J, Zisserman A (2003) Video google: a text retrieval approach to object matching in videos. In: Computer Vision, 2003. Proceedings. Ninth IEEE International Conference on, IEEE, pp 1470–1477

Smolensky P (1986) Information processing in dynamical systems: Foundations of harmony theory. Parallel Distributed Processing: Explorations in the Microstructure of Cognition 1:194–281

Sun Y, Wang X, Tang X (2013) Hybrid deep learning for face verification. In: Computer Vision (ICCV), 2013 IEEE International Conference on, IEEE, pp 1489–1496

Taylor G, Fergus R, LeCun Y, Bregler C (2010) Convolutional learning of spatio-temporal features. In: Computer VisionCECCV 2010, Springer, pp 140–153

Tipping M (2001) Sparse bayesian learning and the relevance vector machine. The Journal of Machine Learning Research 1:211–244

Tsochantaridis I, Hofmann T, Joachims T, Altun Y (2004) Support vector learning for interdependent and structured output spaces. In: Proceedings of the twenty-first international conference on Machine Learning, ACM, p 104

Turaga P, Chellappa R, Subrahmanian V, Udrea O (2008) Machine recognition of human activities: a survey. Circuits and Systems for

Video Technology, IEEE Transactions on 18(11):1473–1488
Wang H, Klaser A, Schmid C, Liu C (2011) Action recognition by dense trajectories. In: Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on, IEEE, pp 3169–3176
Wang Y, Mori G (2009) Human action recognition by semilatent topic models. Pattern Analysis and Machine Intelligence, IEEE Transactions on 31(10):1762–1774
Wei X, Jiang Y, Ngo C (2011) Concept-driven multi-modality fusion for video search. Circuits and Systems for Video Technology, IEEE Transactions on 21(1):62–73
Yang Y, Shah M (2012) Complex events detection using data-driven concepts. In: Computer VisionCECCV 2012, Springer, pp 722–735
Zhao F, Huang Y, Wang L, Tan T (2013) Relevance topic model for unstructured social group activity recognition. In: Advances in Neural Information Processing Systems, MIT Press, pp 2580–2588
Zhu J, Ahmed A, Xing E (2012) Medlda: Maximum margin supervised topic models. The Journal of Machine Learning Research 13(1):2237–2278