Recognizing Actions from Different Views by Topic Transfer

Jia Liu
Department of Electronic Technology, Engineering University of CAPF
Xi'an,Shanxi,China
[e-mail: liujia1022@gmail.com]
*Corresponding author: Jia Liu

Received August 23, 2016; revised December 18, 2016; revised February 13, 2017; accepted February 17, 2017; published April 30, 2017

Abstract

In this paper, we describe a novel method for recognizing human actions from different views via view knowledge transfer. Our approach is characterized by two aspects: 1) We propose a unsupervised topic transfer model (TTM) to model two view-dependent vocabularies, where the original bag of visual words (BoVW) representation can be transferred into a bag of topics (BoT) representation. The higher-level BoT features, which can be shared across views, can connect action models for different views. 2) Our features make it possible to obtain a discriminative model of action under one view and categorize actions in another view. We tested our approach on the IXMAS data set, and the results are promising, given such a simple approach. In addition, we also demonstrate a supervised topic transfer model (STTM), which can combine transfer feature learning and discriminative classifier learning into one framework.

Keywords: action recognition, topic model, transfer learning, cross-view

This research was supported by the National Nature and Science Foundation of China (grant no. 61403417).

https://doi.org/10.3837/tiis.2017.04.015 **ISSN : 1976-7277**

1. Introduction

Human action recognition has received increasing attention in computer vision and plays an important role in practical applications, such as video indexing and retrieval, human-computer interaction, video surveillance, etc.

There are many methods of human action recognition [1] [2] [3] [28], which are extremely powerful in recognizing actions from similar views; however, their performance tend to dramatically decrease when the viewpoint changes. This is primarily because most action representation constructions are based on spatio-temporal patterns of appearance. These low-level features become less discriminative when the action is observed from different views. One possible solution is to train a separate classifier for each viewpoint. However, it is impractical to obtain sufficient examples of each action for each view. In this paper, we argue that higher action knowledge can be transferred across different views by exploring the relationship between low-level, view-dependent features.

Our method can be viewed as a two-stage framework. 1) High-level features, which can be shared across two different views, are learnt. For this purpose, we require types of shared activities that are observed in both source and target views. Firstly, we construct individual visual vocabularies for both views; all of the action videos can be represented as a bag of visual words (BoVWs) [30] [31], in which each frame corresponds to a word. Then, a source topic model for the source view can be constructed using the variational inference method. Each word in the video from the source view can be assigned a topic denoting the higher-level features. Finally, we can now force the corresponding words in the target view to be assigned the same topic as in the source view. A target topic model is trained with these observed words and transferred topics. With the source and target topic models, all of the videos can be represented by a bag of topics (BoTs). These BoTs can transfer activity models from the source to the target view. The new action representations are view invariant. 2) After obtaining the view-invariant features, a classifier that was trained in the source view can be used to recognize actions in the target view. Here, we call the training data orphan activities, which are observed only in the source view. The core idea here is that the topics are transferable because we taught them to be, which means that the bag of topics features of a video in the source view are similar to the topic-based features in the target view.

2. Related Work

There are several approaches that have been proposed for multi-view action recognition. There are three primary categories of approaches used in the literature. One is to use geometry constraints to capture the dramatic changes in actions in different views. Parameswaran et al. [4] defined a view-invariant representation of actions based on the theory of 2D and 3D invariants. The authors considered an action to be a sequence of poses and assumed that there exists at least one key pose in the sequence. Using this assumption, the authors derived a set of view-invariant descriptors. Weinland [5] proposed a framework where actions are characterized by three-dimensional occupancy grids from multiple viewpoints. The researchers in [6] performed 3D reconstruction for multi-view action recognition. Li et al. [7] attempted to estimate 3D shapes and poses from multi-view inputs for action recognition. Liu et al. [29] proposed a 3D action representation which formed as a result of feeding the hierarchical combination of RTs to the Bag of Visual Words model (BoVW) . Weinland et al.

[8] handled viewpoint changes by teaching a hierarchical classifier on a histogram of oriented gradients (HOG) descriptor for training examples taken from various views. The second category of approaches is based on designing features that are well behaved upon a change in domain. Junejo et al. [9] observed that the temporal self-similarity matrices of an action observed from different viewpoints are extremely similar. The authors described a sequence as a histogram of local descriptors, which is calculated from the self-similarity matrix. Farhadi et al. [10] modeled the view as a latent parameter and taught features that can discriminate between views and actions. Their method requires good parameter initialization. The third category is based on learning a transfer feature across different views. Farhadi et al. [11] employed a clustering method to generate split-based features in the source view; then, a support vector machine (SVM) classifier was trained to predict split-based features in the target view. The split-based features are transferable across views. A similar approach was taken by Liu et al. [12]; the authors employed a bipartite graph to model two view-dependent vocabularies and transferred a BoVW action model into a bag-of-bilingual-words (BoBW) model, which is more discriminative in the presence of view changes. This category of method is relevant to transfer learning, which has been explored in machine learning to transfer knowledge across different domains or tasks. An extensive literature review is available in [13]. Recently, knowledge transfer-based methods [11] [12] [22] [23] [24] [25] [27] have become popular for cross-view action recognition. These methods find a view independent latent space in which features extracted from different views are directly comparable. Gupta [26] directly matching purely motion based features from videos to mocap and recovers 3D pose sequences without performing any body part tracking. Rahmani [27] propose unsupervised learning of a non-linear model that transfers knowledge from multiple views to a canonical view. The proposed Non-linear Knowledge Transfer Model (NKTM) is a deep network, with weight decay and sparsity constraints, which finds a shared high-level virtual path from videos captured from different unknown viewpoints to the same canonical view. Notably, our method is similar to those of [11] and [12]; however, there are several significant differences. First, their transfer feature is provided by a trained predictor [11] or a co-cluster [12]; our method, which maps visual words to topics, is straightforward and efficient. Second, it is easier for our method to be extended to a supervised style. The transfer feature and classifier learning can be simultaneously performed.

Our method is also motivated by latent topic models [20] that have been successfully explored in object categories. Niebles et al. [14] used LDA and pLSA for human action categories and location. Wong et al. [15] extended probabilistic latent semantic analysis (pLSA) to capture both semantic (content of parts) and structural (connection between parts) information for motion category recognition. Zhang et al. [16] proposed a new approach, structural pLSA (SpLSA), to explicitly model word orders by introducing latent variables for human action categorization. Wang et al. [17] presented two semi-latent hierarchical topic models for action recognition based on motion words. Bian [18] proposed a transfer topic model that uses information extracted from videos in the auxiliary domain to assist recognition tasks in the target domain.

3. Our Approach

3.1 Low-level Representation

We first describe the low-level action representation method. Similar to [11], we describe frames by vector quantization of local appearance features to form codewords. We use the

feature extraction method from [19]. Their descriptor is a histogram of the silhouette and of the optic flow inside the normalized bounding box. They scale the larger side of the bounding box to a fixed size. Features consist of three channels, horizontal flow, vertical flow and the silhouette. In each channel, the measurements are resampled to fit into the normalized (120×120) box while maintaining the aspect ratio. The normalized feature bounding box is divided into 2×2 sub-windows. Each sub-window is divided into 18-bin radial histogram covering 20 degrees each. The center of the pie is in the center of the sub-window, and the slices do not overlap. The values of each channel are integrated over the domain of every slice. The result is a 72-dimensional histogram. Each of the three channels is separately integrated over the domain of each bin. By concatenating the histograms of all three channels, we obtain a 216-dimensional frame descriptor. To encode the local temporal structure of the actions, we consider 15 frames around the current frame and split them into 3 blocks of 5 frames: past, current and future. The frame descriptors of each block are stacked together into a 1080-dimensional vector which is 5*216-dimension for one block. We then choose the first 50 principal components of the descriptors of a window of size 10 centered at the frame that we want to describe. We keep the first 50, 10 and 10 dimensions for the current, past and future blocks, respectively, which results in 70 dimensional descriptors for each frame, which we call descriptive features. **Fig. 1** depicts the feature extraction procedure. The temporal context descriptor is appended to the current frame descriptor to form the final 286-dimensional motion context descriptor.

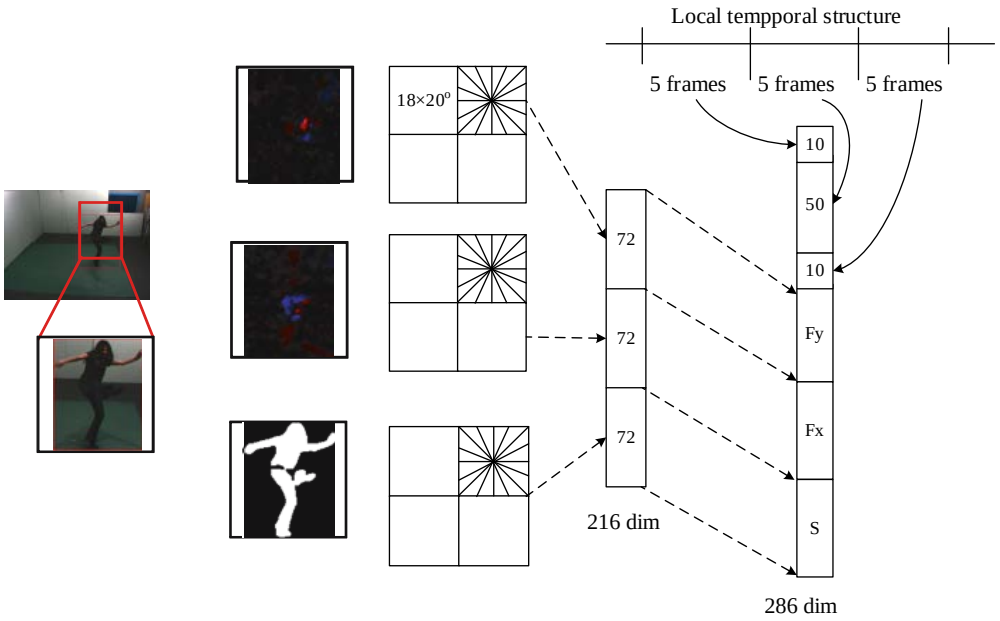


Fig. 1. Feature extraction method.

Our topic transfer models rely on the existence of a finite vocabulary of visual words of size V. To learn the vocabulary of visual words, we consider the set of descriptors corresponding to all frame descriptors in the training data. This vocabulary (or codebook) is constructed by clustering using the K-means algorithm, where the Euclidean distance is used as the clustering metric. The center of each resulting cluster is defined as a codeword; each frame descriptor of an image sequence is then assigned to a visual word.

3.2 Topic Transfer Model

Human actions appear different from different viewpoints. To be able to transfer action knowledge from one view to another view, we require discriminative features that tend to be similar in different views. The bag-of-words feature has been proven to contain discriminative information in similar views. However, this word-based representation is still a low-level feature. We cannot obtain the corresponding relationship between source view words and target view words. We require an approach that explores the corresponding relationships. In this paper, we use the topic model, latent Dirichlet allocation [20], to represent the higher-level features, which can be transferred between different views. In this section, we first describe the "Bag of Topics" representation; then, we present the details of our topic transfer framework. A cross-view action classifier is trained with these higher features for action recognition.

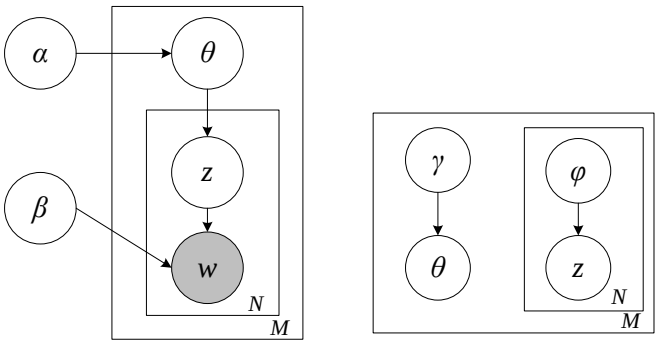
Bag of Topics Representation

Suppose we are given a collection of M video sequences $\{\mathbf{w_1},\mathbf{w_2}, ...,\mathbf{w_M}\}$. Each video sequence $\mathbf{w}$ is a collection of words $\mathbf{w} = (w_1,w_2, ...,w_N)$, where w_i is a codeword for a frame descriptor. A word is a basic item from a codebook indexed by $\{1,2,\ldots V_m\}$,where V_m is the size of the codebook . We exploit the 1-of-K coded binary vector representation, which has been widely used for multiclass classifier designs.

The LDA model assumes there are K underlying topics according to which video sequences are generated. Each topic is represented by a multinomial distribution over the V codewords. A video sequence is generated by sampling a mixture of these topics followed by sampling motion words conditioning on particular topics. These topics describe the higher semantic relations between documents and word terms.

The generative process of LDA for video sequence $\mathbf{w}$ in the collection can be formalized as follows (see **Fig. 2**):

1) Choose $\theta \sim \mathrm{Dir}(\alpha)$
2) For each of the N codewords w_n:
   a) Choose a topic $z_n \sim \mathrm{Mult}(\theta)$;
   b) Choose a word w_n from $w_n \sim p(w_n|z_n, \beta)$, which is a multinomial probability conditioned on z_n.



(a) (b)

Fig. 2. (a) Latent Dirichlet allocation (LDA) graphical model. (b) Graphical model that represents the variational distributions proposed in [20] to approximate the posterior probability in LDA.

Parameter θ indicates the mixing proportion of different topics in a particular video sequence. α is the parameter of a Dirichlet distribution that controls how the mixing proportion

θ varies among different video sequences. In addition, matrix β of size K ×V parameterizes the distribution of spatial-temporal words conditioned on each topic; each element of β corresponds to the probability $p(w_i|z_k)$, which indicates the distribution of motion words within a particular topic. The probability of a video w = $\{w_1,w_2, ...,w_n\}$ is

$$p(\mathbf{w} \mid \alpha, \beta) = \int p(\theta \mid \alpha) \left(\prod_{n=1}^{N} \sum_{z_n} p(z_n \mid \theta) p(w_n \mid z_n, \beta) \right) d\theta \quad (1)$$

Given a collection of video clips $D = \{\mathbf{w}_1,\mathbf{w}_2, ...,\mathbf{w}_M\}$, teaching an LDA model means estimating the model parameters α and β that maximize the log likelihood of all the video data. This parameter estimation problem can be solved by the variational EM algorithm developed in [20].

Once the parameter β is learnt from the training data, these parameters provide means to depict the codewords in a higher level. Specifically, each word is assigned the highest likelihood given a topic, which means that video representation $\mathbf{w} = (w_1,w_2, ...,w_N)$ can be converted to a topic-based representation, $\mathbf{z} = (z_1,z_2, ...,z_N)$, where topic index $z_i \in \{0, 1\}^K$ is an indicator vector with only one entry of 1 and the rest are all 0. We can now describe actions by a *K*-dimension histogram of the topics; we call this **z** the "bag of topics" representation. Because we represent a frame in an image sequence as a "single topic", we could build topic models for both the source and the target views. However, this is not enough because the LDA is an unsupervised generative model, and we still do not know which topic in the target view corresponds to which one in the source view; thus, we could not transfer a model.

Topic Transfer Framework

To make the "bag of topics" representation insensitive to view change, we must teach a topic transfer model. **Fig. 3** depicts the entire procedure for constructing the transfer bag-of-topics features. This procedure can be divided into four steps as follows:

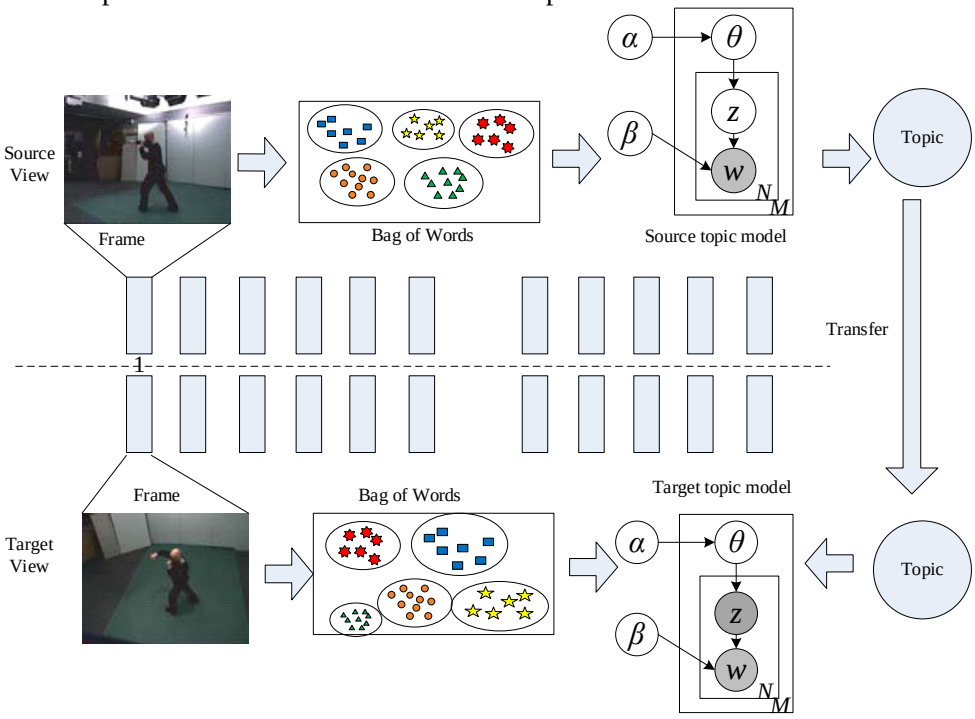


Fig. 3. Chart flow of topic transfer model for cross-view representation.

(1) For both source and target views, we begin by clustering all the frame descriptors to form the codewords. One video in the source view and its corresponding video in the target view can be described by the bag of words, $\mathbf{w_s} = (w_{s1}, w_{s2}, ..., w_{sN})$ and $\mathbf{w_t} = (w_{t1}, w_{t2}, ..., w_{tN})$, respectively.

(2) In the source view, we can use a source topic model, such as LDA, to obtain a bag-of-topic representation for video $\mathbf{w_s}$ .Our topic-based features take the form $\mathbf{z_s} = (z_{s1}, z_{s2}, ..., z_{sN})$.

(3) We can directly transfer the topic-based features to the corresponding frames in the target view. Using unlabeled shared activities, which have established implicit correspondences, topics are transferred from the source to the target view. The topic-based features in the target view take the form $\mathbf{z_t} = (z_{t1}, z_{t2}, ..., z_{tN})$, where $z_{ti} = z_{si}$.

(4) With the word-based features $\mathbf{w_t}$ and the topic-based features $\mathbf{z_t}$, we now can teach a target topic model in the target view. This procedure is depicted in **Fig. 4**.

Notably, **Fig. 4** is not a standard probabilistic graphical model. Here we introduce a new graphical representation for the topic transfer model, in which we add a dashed line with an arrow from the source topic model to the target topic model. This dashed link does not express probabilistic relationships between variables z^s and z^t, where z^s and z^t are the topics in source domain and target domain. This dashed link not only denotes the transfer direction, but this also means that the variables z^s and z^t are equal. The most important aspect is that the topic variables z^s in the source topic model are latent variables; we must use the approximate inference method to obtain its value. This value is directly transferred to the target view. Therefore, in the target view, the topic variable can be viewed as an observed variable. This type of topic model is called a semi-latent topic model. Wang [17] first proposed two semi-latent hierarchical topic models for action recognition based on motion words. In our work, the target topic model is a semi-latent Dirichlet allocation (semi-LDA), where the maximum-likelihood estimate of β can be calculated by simply counting the frequency of each word that appear together with topic z_i, $\beta_{ij}=n_{ij}/ n_{i,}$, where n_i is the count of the i-th topic in the corpus, and n_{ij} is the count of the i-th topic with the j-th word in the corpus. The Dirichlet parameter α can be estimated from a "Dirichlet-multinomial" distribution. We should emphasize that the target topic model in **Fig. 4** is only for training. In testing, we will use the same model as the LDA to infer the topic z^t together with the estimated model parameters α^t *and* β^t.

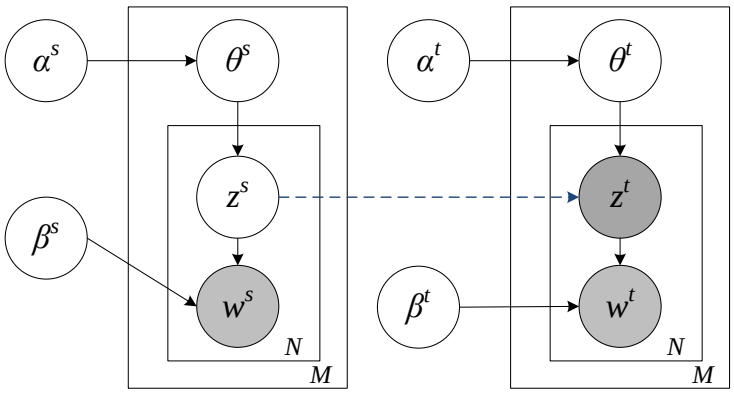


Fig. 4. Graphical representation for the topic transfer model. On the left-hand side, the source topic model is shown, and on the right-hand side, the topic model for the target view is shown.

Thus far, we have the *source topic model*, *target topic model* and the topic-based features in both views. For shared activities, the topic-based features are the same in both views. With this topic transfer model, we construct a semantically similar feature space for different views. This topic-based feature can be viewed as the higher-level features that can be shared across views and can connect action models for different views.

Action Recognition

With the topic transfer model, we can describe each frame using the topic-based features in both views. With enough training data, we will have similar topic-based features for orphan activities in both views. This means that we can simply use the classifier taught on the source view and test it in the target view. In this paper, we use a multi-class SVM as the classifier. **Table 1**. gives the general framework to build topic transferable models.

Table 1. General framework for topic transfer models

General framework for topic transfer models
Learning transferable features: 1. Extract the bag-of-words feature of shared activities in source view and target view. 2. Teach a source topic model using LDA in the source view. 3. Transfer the topics from source view to target view. 4. Teach a target topic model using semi-LDA in the target view.
Cross-view action recognition: 1. Training a classifier: a) Extract bag-of-words features for orphan activities in the source view. b) Use the source topic model LDA to construct the topic-based features. c) Teach a multi-class classifier with SVM and topic-based features. 2. Recognize the query action from target view. a) Extract the bag-of-words features for query clips in the target view. b) Construct the topic-based features in the target view using the already taught target topic model semi-LDA in the target view. c) Recognize the query action clip using the multi-class classifier that was taught in the source view.

3.3 Supervised Topic Transfer Model

In the above transfer model, we assume that the shared activities, which were used to learn the transfer model, are unlabeled. We use an unsupervised topic model LDA to represent the higher-level features. We also need to train a multi-classifier to recognize the query action clip. Under most conditions, all the frames in the shared activities video sequences have action class labels associated with them. In this case, there is no reason to ignore this important information. In this section, we introduce a supervised form of a previous topic transfer model,

called the supervised topic transfer model (STTM). STTM uses class labels by constructing a supervised topic model for the source view. The graphical representation of the STTM model is shown in **Fig. 5**.

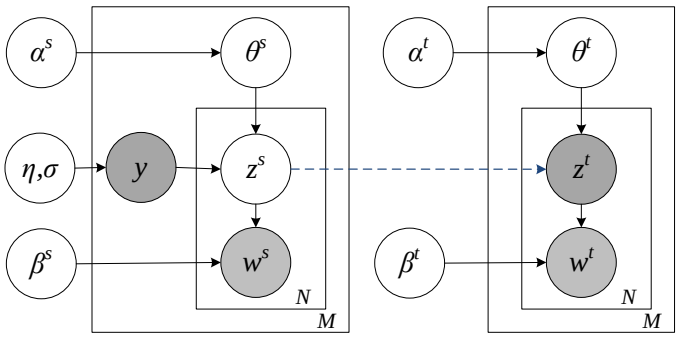


Fig. 5. Graphical representation for the supervised topic transfer model. The left-hand side shows the source topic model, which uses class labels, and the right-hand side shows the semi-LDA used for the target view.

This supervised model also contains two parts. One part is a supervised LDA, which was proposed in Blei et al. [21].The goal of the supervised LDA is to infer latent topics predictive of the response. Given unlabeled data, a supervised LDA can not only infer its topic structure using a fitted model but can also form its label prediction. Our goal is to transfer view knowledge across views such that the supervised topic model trained in the source view can be used to recognize novel actions taken in the target view. The other part of the supervised model is still a semi-LDA.

Suppose a K-topic model in source view, where K topics are denoted by $\beta = \{\beta_1,\beta_2,...,\beta_K\}$; video **v** is associated with a pair (**w**, y), where **w** = $\{w_1,w_2,...,w_N\}$ is its bag-of-words representation and y is its class label; then, the following is the generation procedure:

1) Sample topic proportions θ from Dirichlet distribution Dir($\theta|\alpha$);
2) For each of N words w_n
   a) sample a topic z_n from multinomial distribution Multi(z_n |θ);
   b) sample a word w_n from multinomial distribution Multi(w_n |β_{zn});
3) Draw response variable y from GLM $(\bar{z}, \eta, \delta)$, where GLM denotes the generalized linear model and $\bar{z} = (1/N)\sum_{n=1}^{N} z_n$.

This parameter estimation problem can be solved by the variational EM algorithm developed by Blei et al. [21]. The supervised topic transfer model has a major advantage over previous approaches. This model combines the topic-based feature learning and classifier learning into one process. Therefore, there is no need to use the orphan activities to train a classifier. Under the supervised framework, a supervised LDA is first trained by the shared activities in the source view; then, the topics are transferred to the corresponding frames in target view. A semi-LDA is taught in the target view. For a new query video clip, we can use the semi-LDA to obtain its topic-based feature; this transferable feature can be delivered to the source supervised-LDA, and the distribution of the class label is a generalized linear model:

$$p(y \mid z_{1:N}^{t}, \eta, \delta) = h(y, \delta) \exp\left(\frac{(\eta^T \bar{z}) y - A(\eta^T \bar{z})}{\delta} \right). \quad (2)$$

where $h(y,\delta)=(1/\sqrt{2\pi\delta})\ \exp\{-y^2/(2\delta)\}$, A is the $D\times(K+1)$ matrix whose rows are the vectors $\bar{z}^{T}$, η and δ are parameters of the supervised LDA, and $z_{1:N}^{t}$ is the topic-based feature of the query clip in target view.

4. Experimental

We tested our approach on the IXMAS multi-view action database [5] (see **Fig. 6** for several examples), which contains 11 daily-live actions. Each action is performed three times by 12 actors taken from five different views: four side views and one top view.

To better compare with the results from [11] [12], we attempted all 20 combinations of transfers among views (there are five views in the IXMAS dataset). First, we clustered the shared activities to 1000 clusters using k-means clustering on the descriptive features to form a codebook. With these basic view-dependent vocabularies, we set the topic number K=100 for both the source and target topic model and taught an LDA model in the source view and transferred the topic to the target view. Actually, we set number of codebook C = {500, 1000,2000}, and topic number K = {50, 100,200},the best result achieved at C =1000, K=100.

A semi-LDA was trained in the target view. For a given orphan action, we first predicted the topic item of each word and then constructed the topic-based description of that action by a histogram of topics. Under the unsupervised topic transfer model, we follow the "leave-one-action-class-out strategy" in [11], which means that each time, we only consider one action class in testing in the target view (this action class is not used to teach the feature transfer model). The final results are reported in terms of the average accuracy for all action classes in each view. The shared activities training data used for constructing a topic transfer model are randomly selected from actions, excluding the orphan action. With the learnt topic transfer model, multi-class SVM classifiers are trained in the source view and used to recognize actions from the target view. In this paper, we used the histogram intersection kernel as the kernel function. Under the supervised topic transfer model, training data used for constructing the supervised topic transfer model are randomly selected from each type of action.

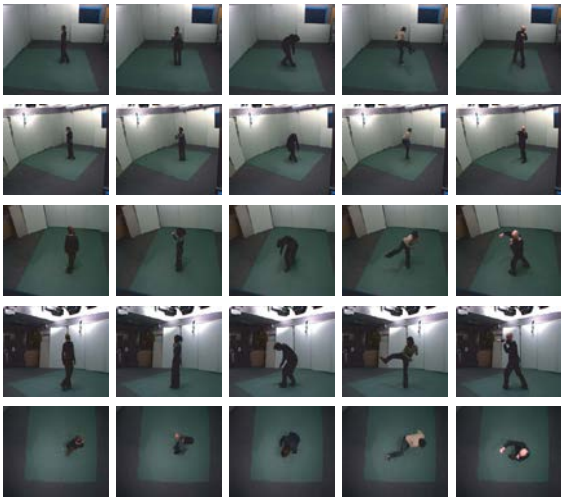


Fig. 6. Example frames from the IXMAS dataset. Each row gives one action from different views.

First, we want to verify the performance of transferring models across pairwise views. Initially, each query action video is represented by a bag-of-words model. We first attempt to recognize novel actions from the target view by directly using classifiers trained on the source view without model transfer. This means that we directly teach a classifier on the bag-of-words feature for the source view and then test it in the target view. The results are shown in **Table 2** (i.e., the woTr columns). This result shows that word-based description works poorly under transfer circumstances. Therefore, we transferred all actions in both views from the "bag of words" to "bag of topics" representation with the topic transfer model. The number of topics is 100 for this experiment. The results are shown in **Table 2** (i.e., the wTr columns).

In **Table 2,** the rows and columns correspond to the training and testing views, respectively. The woTr columns and wTr columns contain the results of recognition with and without topic model transfer. The average accuracies are **10.2%** and **72.7%** for woTran and wTran, respectively. Considering that the classifiers are trained based on data taken from different views, the performance is extremely promising. The result also demonstrates that "bag of topics" representation is discriminative under view changes.

Table 2. Performance comparison of action recognition with and without model transfer.

(%)	**Cam0**		**Cam1**		**Cam2**		**Cam3**		**Cam4**	
	woTr	wTr	woTr	wTr	woTr	wTr	woTr	wTr	woTr	wTr
Cam0	---	---	13.5	66.8	10.8	67.4	8.2	65.1	10.5	68.4
Cam1	14.5	67.4	---	---	9.3	63.9	6.8	68.4	8.5	71.5
Cam2	12.3	**88.7**	9.8	**85.2**	---	---	7.8	**86.9**	9.6	**89.9**
Cam3	16.7	67.4	10.2	68.6	8.7	68.0	---	---	7.4	65.1
Cam4	11.5	73.7	7.6	73.3	6.9	73.9	11.4	75.1	---	---

To further demonstrate the performance of cross-view action recognition, we applied classification in single view, which means we trained and tested the classifiers on the same view. The topic model provides a dimension reduction method. The "bag of topics representation" is able to capture semantic relationships between word topics and topic documents interpreted in terms of probability distributions. The results are shown in Table 3. The SwT columns and SwoT columns contain the results of recognition with and without topic model in single view. Notably, the performance of cross-view recognition shown in Table 2 is extremely close to the single-view classification result shown in **Table 3.**

Table 3. Performance comparison of action recognition with and without topic representation in single view.

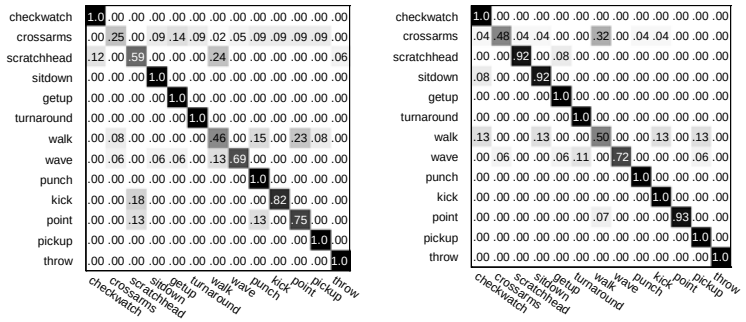
View	**Cam0**		**Cam1**		**Cam2**		**Cam3**		**Cam4**	
(%)	SwoT	SwT	SwoT	SwT	SwoT	SwT	SwoT	SwT	SwoT	SwT
	74.2	76.2	78.4	77.5	75.2	72.6	65.2	66.2	71.6	70.4

We also give two additional sets of state-of-the-art results on cross-view action recognition reported in [11] and [12] in Table 4 for comparison. For a better comparison to these method, we use the unsupervised topic model for cross-view action recognition. Columns O, F and L represent our approach, Farhadi's approach [11] and Liu's approach [12], respectively. We are particularly interested in [11] because they used the same frame-to-frame correspondence relationship as ours to train a model across views. Our results are competitive to Farhadi's methods. The performance of our model is lower than that of [12]. In their method, Liu uses two types of features; one is the 3D interest point feature, and the other is the shape-flow descriptor, which is the same as ours. The performance has been improved in terms of average accuracy by combining the shape-flow descriptor and interest point feature.

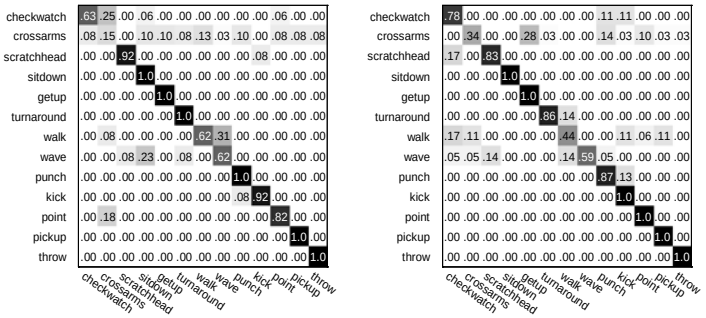
Table 4. Cross-view action recognition performance of different approaches on the IXMAS dataset.

%	**Cam0**			**Cam1**			**Cam2**			**Cam3**			**Cam4**		
	F	L	O	F	L	O	F	L	O	F	L	O	F	L	O
C0	---	---	---	72	79	66	61	76	67	62	76	65	30	74	68
C1	69	81	67	---	---	---	64	75	63	68	78	68	41	70	**71**
C2	62	79	**88**	67	76	**85**	---	---	---	67	79	**86**	43	72	**89**
C3	63	73	67	72	74	68	68	74	**68**	---	---	---	44	66	65
C4	51	82	73	55	68	**73**	51	74	73	53	71	**75**	---	---	---
AV	61	79	74	67	74	73	61	75	67	62	76	73	40	71	**73**

We further checked the confusion matrices generated from our topic transfer model. **Fig. 7** shows the confusion matrices for camera 1. In this experiment, the video from target view camera 1 are the test data. Data from cameras 0, 2, 3 and 4 are used to train a classifier. As seen in **Fig. 7**, several actions, such as "cross arms" and "wave" are difficult to distinguish when observed from a certain viewpoint. This is because it is difficult to transfer the topic transfer models from various views to a certain view for these actions.



(a) TrainCam0TestCam1 (b) TrainCam2TestCam1



(a) TrainCam3TestCam1 (b) TrainCam4TestCam1

Fig. 7. Confusion matrices for camera 1.

Additionally, we want to demonstrate the performance of the supervised topic model. The GLM framework gives us the ability to model any type of response variable whose distribution can be written in exponential dispersion. In this experiment, we use the multinomial distribution as the response distribution. As seen in **Table 5**, the supervised topic transfer model achieved a competitive result to the original topic transfer, where the performance is slightly lower than the unsupervised topic transfer with multi-class SVM. This may be due because discriminative training might give better results than direct use of the generative model.

Table 5. Cross-view action recognition performance of the supervised topic transfer model

(%)	**Cam0**	**Cam1**	**Cam2**	**Cam3**	**Cam4**
Cam0	----	65	62	60	62
Cam1	68	----	61	61	61
Cam2	75	72	----	71	60
Cam3	60	64	63	----	63
Cam4	63	64	62	63	----

Table 6. The cross-view recognition results of different cross-view approaches

Approach	Accuracy(%)
Ours	64.0%
Hankelets[25]	56.4%
nCTE[26]	67.4%
NKTM[27]	72.5%
CVP[23]	69.02%
VISP[24]	98.02%

We also list additional sets of state-of-the-art results on cross-view action recognition reported in [25] and [26], [27], [23], [24] in **Table 6**. Our method is slightly lower than

nCTE[26] and NKTM[27] , However, nCTE[26] requires 30GB memory to store mocap samples. NKTM[27] using dense trajectories which extracted from videos, this costs long time for feature extraction. Our method, which maps visual words to topics, is straightforward and efficient. In the classification problem, discriminative approach such as [23] and [24] are superior. Our method is based on topic model, which is a generative model estimate the joint probability density function, still have a number of attractive properties. First, our method is an intuitive solution to knowledge transfer. Second, the prior knowledge can be easier to integrate into a graphical topic model. It is note that ref. [24] not only learns a common dictionary that models the view-shared features, but also learns a dictionary pair corresponding to the source and target views to model view-specific features. Our approach enforces the topic zs in source view are equal to the zt in the target view, view-shared features (topics) and view-specific features (BOVWs) in our approach are more intuitive and simple.

In the above experiments, we assumed that the view of action is known. This means that we know the source and the target view. When the target view is unavailable, we can train a SVM classifier to detect the view discriminatively using the low-level features described in section 3.1. **Table 4** gives the classification results for the view classification problem. If the target view is not specified, we should multiply accuracies in **Table 4** by the accuracies in **Table 7.** Because the numbers in **Table 7** are extremely close to 1, we should not expect a major change in accuracies. Notably, our view classification method is similar to that by Farhadi [11]. The difference is that we use the SVM for classification and is not a 1-nearest-neighbor (1NN) classifier.

Table 7. Classification results for view classification problem.

Cam0	Cam1	Cam2	Cam3	Cam4
95%	92%	97%	94%	83%

5. Conclusion

In this paper, we proposed a topic transfer model for cross-view action recognition. We constructed a semantically similar feature space for different views by teaching source and target topic models. With these models, the original bag of visual words (BoVW) representation can be transferred into a bag of topics (BoT) representation. These topic-based features can be seen as view-invariant features. Therefore, the discriminative classifier of actions can be trained in source view and tested in a different target view. We extensively tested our approach using the publicly available IXMAS multi-view data set, and the experiment results are competitive with the best results reported in the literature. Additionally, to use the action class labels in shared activates, we proposed a supervised topic transfer model that combines transfer feature learning and discriminative classifier learning in a single procedure.

References

[1] Laptev, I., "On space-time interest points," *International Journal of Computer Vision*, 64(2-3):pp. 107-123, 2005. Article (CrossRef Link).

[2] Dollar, P.R., V. Cottrell, and G. Belongie, S., "Behavior recognition via sparse spatio-temporal features," in *Proc. of IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance*, pp. 1-8. 2005. Article (CrossRef Link).

[3] Fengjun, L and Nevatia, R., "Single view human action recognition using key pose matching and Viterbi path searching," in *Proc. of IEEE Conference on Computer Vision and Pattern Recognition*, p. 1304-1311, 2007. Article (CrossRef Link).

[4] Parameswaran, V.C., R., "View invariance for human action recognition," I*nternational Journal of Computer Vision*, 66(1): p. 83-101, 2006. Article (CrossRef Link).

[5] Weinland, D., E. Boyer, and R. Ronfard, "Action recognition from arbitrary views using 3D exemplars," in *Proc. of IEEE International Conference on Computer Vision*, pp. 170-176. 2007. Article (CrossRef Link).

[6] Pingkun, Y.K., S. M. and Shah, M., "Learning 4D action feature models for arbitrary view action recognition," in *Proc. of IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7-15, 2008. Article (CrossRef Link).

[7] R. Li, T.T., and S. Sclaroff, "Simultaneous learning of nonlinear manifold and dynamical models for high-dimensional time series," in *Proc. of International Conference of Computer Vision*, pp. 1-8, 2007. Article (CrossRef Link).

[8] D. Weinland, M. Ozuysal, and P. Fua, "Making action recognition robust to occlusions and viewpoint changes," in *Proc. of Europe Conference on Computer Vision*, pp. 635-648, 2010. Article (CrossRef Link).

[9] IN. Junejo, E Dexter, I. Laptev and P. Pérez, "Cross-View Action Recognition from Temporal Self-similarities," in *Proc. of Europe Conference on Computer Vision*, pp. 293-306. 2008. Article (CrossRef Link).

[10] A. Farhadi, M. Tabrizi, I. Endres, and D. Forsyth, "A latent model of discriminative aspect," in *Proc. of International Conference of Computer Vision*, pp. 948-955, 2009. Article (CrossRef Link).

[11] Farhadi, A.F., D. White, R., "Transfer learning in sign language," in *Proc. of IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2909-2916, 2007. Article (CrossRef Link).

[12] Liu, J.G.S., M. Kuipers and B. Savarese, S., "Cross-View Action Recognition via View Knowledge Transfer," in *Proc. of IEEE Conference on Computer Vision and Pattern Recognition* 2011, IEEE: New York. pp. 3209-3216, 2011. Article (CrossRef Link).

[13] Sinno Jialin, and P.Q., Yang, "A Survey on Transfer Learning," *IEEE Transactions on Knowledge and Data Engineering*, 22(10): pp. 1345-1359, 2010. Article (CrossRef Link).

[14] J.C. Niebles, H-C. Wang and F.F. Li, "Unsupervised learning of human action categories using spatial–temporal words," *International Journal of Computer Vision* 79 (3), pp.299-318, 2008. Article (CrossRef Link).

[15] S-F. Wong, T-K. Kim and R. Cipolla, "Learning motion categories using both semantic and structural information," in *Proc. of IEEE Conference on Computer Vision and Pattern Recognition*, pp.18-23 2007. Article (CrossRef Link).

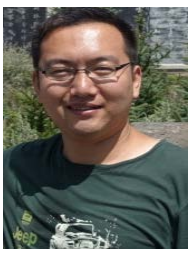
[16] J.G. Zhang and S.H. Gong, "Action categorization by structural probabilistic latent semantic analysis," *Computer Vision and Image Understanding*. 114(8), pp. 857-864, 2010. Article (CrossRef Link).

[17] Y. Wang and G. Mori, "Human Action Recognition by Semi-latent Topic Models," *IEEE Trans. Pattern Anal. Mach. Intell*. 31(10), pp.1762-1774, 2010. Article (CrossRef Link).

[18] Bian, W, Tao, D. C. and Rui, Y., "Cross-Domain Human Action Recognition," *IEEE Transactions on Systems Man and Cybernetics Part B-Cybernetics*, 42(2): pp. 298-307. 2012. Article (CrossRef Link).

[19] Du Tran and Alexander Sorokin, "Human activity recognition with metric learning," in *Proc. of European Conference on Computer Vision*, pp. 548-561, 2008. Article (CrossRef Link).

[20] Blei, D.M. Ng., A. Y. and Jordan, M. I., "Latent Dirichlet allocation," *Journal of Machine Learning Research*, 3(4-5): pp. 993-1022, 2003. Article (CrossRef Link).
[21] David M. Blei and Jon D. McAuliffe, "Supervised topic models," in *Proc. of Advances in Neural Information Processing Systems*, pp. 1-8. 2007. Article (CrossRef Link).
[22] R. Li and T. Zickler, "Discriminative virtual views for cross-view action recognition," in *Proc. of IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1-8. 2012. Article (CrossRef Link).
[23] Z. Zhang, C. Wang, B. Xiao, W. Zhou, S. Liu, and C. Shi, "Cross-view action recognition via a continuous virtual path," in *Proc. of IEEE Conference on Computer Vision and Pattern Recognition*,, pp. 1-8. 2013. Article (CrossRef Link).
[24] J. Zheng and Z. Jiang, "Learning view-invariant sparse representations for cross-view action recognition," in *Proc. of IEEE International Conference on Computer Vision*, pp. 1-8, 2013. Article (CrossRef Link).
[25] B. Li, O. Camps, and M. Sznaier, "Cross-view activity recognition using hankelets," in *Proc. of IEEE International Conference on Computer Vision*, pp. 1-8, 2012. Article (CrossRef Link).
[26] A. Gupta , J. Martinez , J. Little and J. Woodham, "3D Pose from Motion for Cross-view Action Recognition via Non-linear Circulant Temporal Encoding," in *Proc. of IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1-8, 2014. Article (CrossRef Link).
[27] H. Rahmani, A. Mian, "Learning a Non-linear Knowledge Transfer Model for Cross-View Action Recognition," in *Proc. of IEEE International Conference on Computer Vision*, pp. 1-8, 2015. Article (CrossRef Link).
[28] M. Liu, H Liu and Q Sun, "Action classification by exploring directional co-occurrence of weighted STIPS," in *Proc. of IEEE International Conference on Image Processing*. pp. 1460-1464, 2014. Article (CrossRef Link).
[29] M. Liu, H. Liu, C Chen and M Najafian, "Energy-based Global Ternary Image for Action Recognition Using Sole Depth Sequences," in *Proc. of International Conference on 3d Vision* ,pp. 1-5, 2016. Article (CrossRef Link).
[30] C. Chen, R. Jafari and N.Kehtarnavaz., "Action recognition from depth sequences using depth motion maps-based local binary patterns," in *Proc. of the IEEE Winter Conference on Applications of Computer Vision*, pp. 1092-1099, 2015. Article (CrossRef Link).
[31] H. Liu, M. Liu, and Q Sun, "Learning directional cooccurrence for human action classification," in *Proc. of International Conference on Acoustics, Speech and Signal Processing*, pp. 1244-1248, 2014. Article (CrossRef Link).



Jia Liu is assistant professor at Department of Electronic Technology, Engineering University of Armed Police Force, Xi'an, China. He received his BE degree and MS degree from the Engineering College of Armed Police Force, Xi'an, China. He received his PH.D. degree in Shanghai Jiaotong University, China. His research interests are computer vision, machine learning, action recognition.