Optical Engineering 50(2), 027203 (February 2011)

Hybrid generative-discriminative human action recognition by combining spatiotemporal words with supervised topic models

Hao Sun
National University of Defense Technology
School of Electrical Science and Engineering
47 Yanwachi Street
Changsha, Hunan 410073, China
E-mail: clhaosun@gmail.com

Cheng Wang
Boliang Wang
Xiamen University
School of Information Science and Technology
Department of Computer Science
Xiamen, Fujian 361005, China

Abstract. We present a hybrid generative-discriminative learning method for human action recognition from video sequences. Our model combines a bag-of-words component with supervised latent topic models. A video sequence is represented as a collection of spatiotemporal words by extracting space–time interest points and describing these points using both shape and motion cues. The supervised latent Dirichlet allocation (sLDA) topic model, which employs discriminative learning using labeled data under a generative framework, is introduced to discover the latent topic structure that is most relevant to action categorization. The proposed algorithm retains most of the desirable properties of generative learning while increasing the classification performance though a discriminative setting. It has also been extended to exploit both labeled data and unlabeled data to learn human actions under a unified framework. We test our algorithm on three challenging data sets: the KTH human motion data set, the Weizmann human action data set, and a ballet data set. Our results are either comparable to or significantly better than previously published results on these data sets and reflect the promise of hybrid generative-discriminative learning approaches. [DOI: 10.1117/1.3537969]

Subject terms: action categorization; spatiotemporal words; supervised topic models; hybrid generative-discriminative learning.

Paper 100828R received Oct. 13, 2010; revised manuscript received Dec. 4, 2010; accepted for publication Dec. 21, 2010; published online Feb. 22, 2011.

1 Introduction

Identifying human actions in video sequences is an appealing yet challenging problem in computer vision with many applications, including motion capture, human-computer interaction, environment control, video summarization, security surveillance, and sport and entertainment analysis. Various features (e.g., holistic or local motion and shape templates) can be used for recognizing actions. In this paper, we focus on recognizing the action of a person in a video sequence based on local space–time features. We develop a novel model for human action recognition based on the bag-of-words paradigm and topic models.

Action recognition has a long history of research with significant progress reported over the last few years. It remains, however, a challenging problem for computers to achieve robust action recognition due to cluttered backgrounds, camera motion, occlusion, view point changes, and geometric and photometric variances of objects.[1] A cluttered background makes it hard to segment the foreground. Camera motion creates ambiguities in the motion patterns. Human actions can also be observed only partially due to occlusions. Viewpoint changes as well as geometric and photometric variances produce very different appearances and shapes for the same category examples, resulting in high intraclass variances.

An action-recognition problem is basically a classification problem, and there are many different modeling approaches for the solution. These approaches can be classified into two main categories, such as generative and discriminative, each offering important distinct advantages.[2] Let y be the label of the class and x the measured data associated with that class. A generative approach will estimate the joint probability density function $p(x,y)$ [or equivalently $p(x|y)$ and $p(y)$] and will classify using $p(y|x)$, which is obtained using Bayes' rule (Fig. 1). Conversely, discriminative approaches will estimate $p(y|x)$ (or alternatively a classification function $y = f(x)$] directly from the data and do not allow one to generate samples from the joint distribution. Examples of generative models include Gaussian distribution, Gaussian mixture model, hidden Markov model, naïve Bayes, latent topic models, and multinomial distribution. Examples of discriminative models include linear discriminant analysis, support vector machines, boosting, conditional random fields, logistic regression, and neural networks. Discriminative approaches have shown better performance given enough data because they are better tailored to the prediction task and appear more robust to model misspecification. Despite the strong empirical success of discriminative methods in a wide range of applications, when the structures to be learned become more complex than the amount of training data (e.g., in machine translation, scene understanding, activity perception), some other source of information must be used to constrain the space of candidate models (e.g., unlabeled examples, related data sources, or human prior knowledge). Generative modeling is a principled way of encoding this additional information (e.g., through probabilistic graphical models or stochastic grammar rules). Moreover, they provide a natural way to use unlabeled data and are sometimes more computationally efficient.

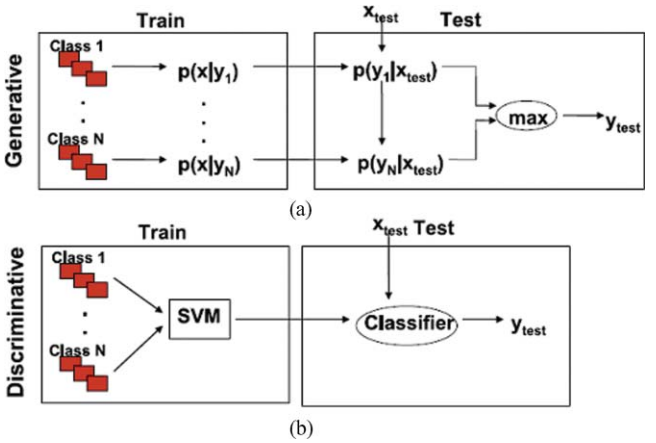


Fig. 1 Schematic comparison of the (a) generative and (b) discriminative approaches for human action categorization.

We propose a hybrid generative-discriminative model approach to learn and recognize human actions in video sequences, taking advantage of the robust representation of the bag-of-words and a supervised generative learning approach. We advocate the use of a hybrid learning setting in an attempt to gain the benefit of both generative and discriminative approaches.

Our approach is motivated by the recent success of the bag-of-words model for general object recognition in computer vision.[3,4] This representation, which is adapted from the text retrieval literature, models the object by the distribution of words from a fixed visual codebook. The common paradigm of these approaches consists of extracting local features from a collection of images, constructing a codebook of visual words by vector quantization, and building a probabilistic model, such as latent Dirichlet allocation (LDA) (Ref. 5) and probabilistic latent semantic indexing (pLSI),[6] to represent the collection of visual words. Probabilistic topic models find a low-dimensional representation of data under the assumption that each data point can exhibit multiple components or topics. In the context of our problem, video is represented as a collection of visual words and our model finds a set of topics that are predictive of action categories in the video sequences.

The rest of the paper is organized as follows. We review previous related work in Sec. 2. In Sec. 3, we describe our approach in more detail, including the bag-of-spatiotemporal feature representation, a brief overview of the supervised LDA model in our context, and the specifics of the learning and recognition procedures. In Sec. 4, we present the experimental results on human action recognition using real data sets, and also compare our performance to other methods. Finally, Sec. 5 concludes the paper.

2 Previous Works

2.1 *Action Recognition*

A considerable amount of previous work has addressed the problem of human action categorization. The approaches in the literature can be broadly categorized as model-based, spatiotemporal template–based, and bag-of-words–based methods.

Model-based approaches first track body parts and then use the obtained motion trajectories to perform action recognition. Ramanan and Forsyth[7] approach action recognition by first tracking humans in the sequences using a pictorial structure procedure. Then 3-D body configurations are estimated and compared to a highly annotated 3-D motion library. In the work by Yilmaz and Shah,[8] human labeling of landmark points in the human body is first done at each frame in sequences from multiple moving cameras. Then, actions are compared using their corresponding trajectories. In the work by Song et al.[9] and Fanti et al.,[10] feature points are first detected and tracked in a frame-by-frame manner. Multiple cues, such as position, velocities, and appearance, are obtained from these tracks. Then, human actions are modeled utilizing graphical models based on triangulated graphs. Spatiotemporal template–based methods analyze human actions by looking at video sequences as space–time intensity volumes. In the work of Bobick and Davis,[11] motion-energy-image and motion-history-image are introduced as templates for different motion recognition. Their method depends on background subtraction and thus cannot tolerate moving cameras and dynamic backgrounds. Blank et al.[12] represent actions as space–time shapes for action recognition. Efros et al.[13] proposed a spatiotemporal descriptor based on global optical flow measurements.

Spatiotemporal template approaches are holistic approaches where global descriptors are used with no local features extracted. In contrast, bag-of-words–based approaches detect local salient features as visual words, which are then used to recognize the activity. Several methods for feature localization and description have been proposed in the literature, and promising recognition results were demonstrated for a number of action classes. Laptev[14] presents a space–time interest point detector based on the idea of the Harris and Forstner interest point operators. He detects local structures in space–time, where the image values have significant local variations in both dimensions. Dollar et al.[15] propose a detector based on a set of separable linear filters, which responds to local regions that exhibit complex motion patterns, including space–time corners. Also, a number of descriptors are proposed for the resulting video patches around each point. Ke et al.[16] apply spatial–temporal volumetric features that efficiently scan video sequences in space and time. Oikonomopoulos et al.[17] extends the idea of saliency regions in spatial images to the spatiotemporal case. The work is based on the spatial interest points of Kadir and Brady.[18]

2.2 *Generative and Discriminative Learning*

Generative and discriminative learning are two of the major paradigms for learning and classifying human actions. When labeled training data are plentiful, discriminative techniques are widely used because they give excellent generalization performance. However, although collection of data is often easy, the process of labeling the data can be expensive. Consequently, there is increasing interest in generative methods because these can exploit unlabeled data in addition to labeled data.

There are many recent studies[19–25] dealing with the comparison of these two approaches to the final goal of combining the two in the best way. In Ref. 19, it was concluded that although the discriminative learning has lower asymptotic error, a generative classifier approaches its higher asymptotic error much faster. In Refs. 20 and 21, discriminative and generative learning were combined in an *ad hoc* manner using a weighting parameter and the value of this parameter

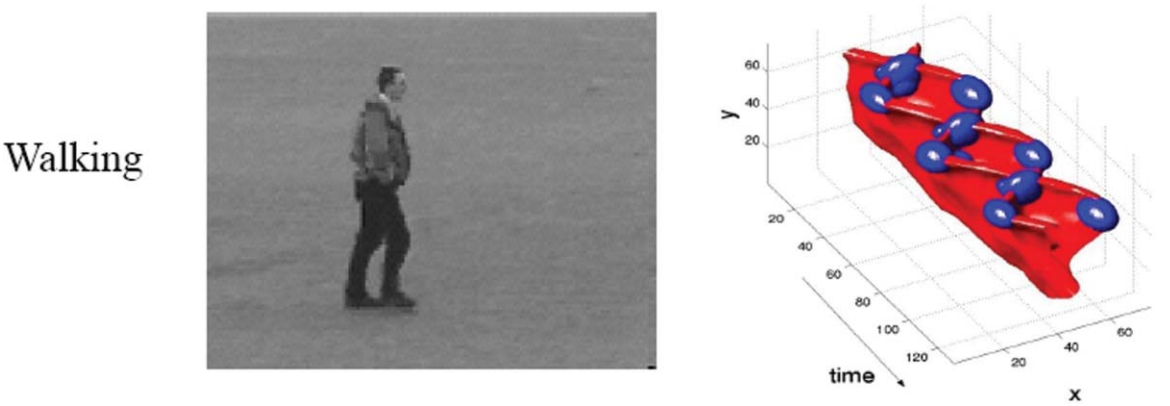


Fig. 2 Human action description using local space–time features.

defines the extent to which discriminative learning is effective over generative learning. In Ref. 22, discriminative learning was performed on a generative model where background posterior probability was modeled with a constant. In Ref. 25, the authors present a method for object category detection that integrates a generative model with a discriminative classifier. For each object category, an appearance codebook, which becomes a common vocabulary for the generative and discriminative methods, is generated.

2.3 *Topic Models for Visual Recognition*

Recently, generative approaches based on latent topic models, such as pLSA and LDA, have been successfully applied in various recognition problems in computer vision. Fei-Fei and Perona[3] use a variant of LDA for natural scene categorization. The words in their model correspond to small local patches in the images, and the topics correspond to the intermediate themes that make up a particular scene. Their model can learn the intermediate themes that are discriminative for different scene categories. Sivic et al.[4] perform unsupervised learning of object categories using variants of the pLSA model. In their models, the words correspond to local patches extracted by interest point operators, and the topics correspond to the different object categories. They later extended their work, experimenting with both pLSA as well as LDA, and using multiple image segments as the equivalent of documents so as to better localize the objects in the images.[26] Fergus et al.[27] extend pLSA to incorporate spatial information in a translation and scale-invariant manner and apply them to learn object categories from Google's image search. Liu and Chen[28] extend the pLSA model with the integration of a temporal model to discover objects in video.

Topic models have also been applied in human activity perception and classification. Bissacco et al.[29] use LDA for the human detection and pose classification. The words in their model are vector quantization of histogram of oriented gradients in the training images. LDA is used to model the intermediate themes that are distinctive for certain human poses. Niebles et al.[1] demonstrate some impressive results on unsupervised learning of human action categories using the pLSA and LDA models. Wong et al.[30] adopt and extend pLSA models to capture both semantic and structural information for recognizing actions and inferring the locations of certain actions. Wang and Mori[31] propose two semi-latent topic models for human action recognition. In their work, each frame from video sequences corresponds to a visual word.

3 Our Approach

3.1 *Spatiotemporal Words for Video Representation*

Local image and video features have been shown successful for many visual recognition problems (e.g., object, scene, and action recognition). Local space–time features capture characteristic shape and motion in video and provide relatively independent representation of events with respect to background clutter and multiple motions in the scene. Such features are usually extracted directly from video and therefore avoid possible failures of other preprocessing methods, such as motion segmentation and tracking. Many different space–time feature detectors and descriptors have been proposed in the past few years.[14–17,32] These space–time features can provide a rich description and powerful representation for human action categorization (Fig. 2).

We use a bag-of-features model and represent each video sequence as a collection of spatiotemporal words by extracting space–time interest points. We detect local space–time features using an extended Harris operator applied at multiple spatial and temporal video resolutions.[14] The Harris 3-D detector computes a spatiotemporal second moment matrix at each video point:

$$\mu(\cdot,\sigma,\tau) = g(\cdot, s\sigma, s\tau) * \{\nabla L(\cdot,\sigma,\tau)[\nabla L(\cdot,\sigma,\tau)]^T\}, \tag{1}$$

where σ and τ are the independent spatial and temporal scale values, g is a separable Gaussian smoothing function, and ∇L computes the space–time gradients. The final locations of space–time interest points are given by local maxima of H,

$$H = \det(\mu) - k\,\mathrm{trace}^3(\mu), \quad H > 0. \tag{2}$$

In our experiments, points are extracted at multiple scales based on a regular sampling of the scale parameters σ and τ. We use parameter settings $k = 0.0005$, $\sigma^2 = 4, 8, 16, 32, 64$, and $\tau^2 = 2.4$. Interest points detected for four consecutive frames of human actions are illustrated in Fig. 3.

For each given sample point (x,y,t,σ,τ), a feature descriptor is computed for a 3-D video patch centered at (x,y,t). Its spatial size $\Delta_x(\sigma)$, $\Delta_y(\sigma)$ is a function of σ and its temporal

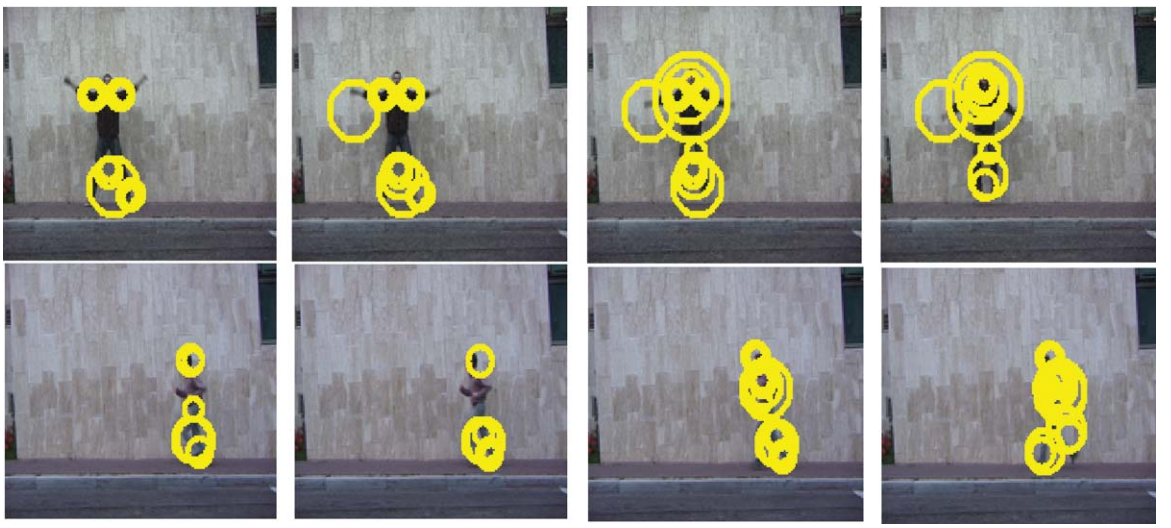


Fig. 3 Space–time Harris interest points detected in human action sequences.

length $\Delta_t(\tau)$ is a function of τ. To characterize local motion and appearance, we compute histograms of oriented gradient (HOG) and histograms of optical flow (HOF) descriptors of space–time volumes accumulated in the neighborhoods of detected interest points. For the combination of HOG and HOF descriptors, the descriptor size is defined by $\Delta_x(\sigma) = \Delta_y(\sigma) = 18\sigma$, $\Delta_t(\tau) = 8\tau$. Each volume is subdivided into a $n_x \times n_y \times n_t$ grid of cells; for each cell, four-bin HOG and five-bin HOG are computed. Normalized histograms are then concatenated to form the final descriptor. In our experiments, we use the grid parameters $n_x, n_y = 3$, $n_t = 2$.

Given a set of spatiotemporal features, we build a spatiotemporal bag of words. This requires the construction of a visual vocabulary. The vocabulary is constructed by clustering using the k-means algorithm and Euclidean distance as the clustering metric. The center of each resulting cluster is defined to be a spatiotemporal word. Thus, each detected point can be assigned a unique cluster membership (i.e., a spatiotemporal word) such that a video can be represented as a collection of spatiotemporal words from the codebook. The effect of the codebook size is explored in our experiments.

3.2 *Learning Actions by Supervised Topic Discovery*

Topic models are distributions over document collections, where each document is represented as a collection of discrete random variables, which are its words. In topic models, we treat the words of a document as arising from a set of latent topics, which are a set of unknown distributions over the vocabulary. Most topic models, such as LDA, are unsupervised. Only the words in the documents are modeled, and the goal is to infer topics that maximize the likelihood or the posterior probability of the collection. Unsupervised LDA has been used to extract latent semantic topics in the collection. However, when the goal is prediction, fitting unsupervised topics may not be a good choice. Consider predicting a movie rating from the words in its review. Intuitively, good predictive topics will differentiate words such as "excellent," "terrible," and "average," without regard to genre. But topics estimated from an unsupervised model may correspond to genres, if that is the dominant structure in the corpus. In the context of our problem, consider predicting human action class from the spatiotemporal words in the video sequences. Similarly, good predictive topics will differentiate words such as "run," "walk," "bend," and "jack" without regard to human poses. But topics estimated from an unsupervised model may correspond to different poses, if that is the dominant structure in the unlabeled training samples.

Unsupervised LDA serves as the basis for many other topic models. In unsupervised LDA, the topic proportion for a document is drawn from a Dirichlet distribution. The words in the document are obtained by repeatedly choosing a topic assignment from those proportions and then drawing a word from the corresponding topic. In supervised LDA (sLDA),[33] a response variable is associated with each document. sLDA jointly models the documents and the responses in order to find latent topics that will best predict the response variables for future unlabeled documents.

Suppose we have a set of $M(j = 1,2,\ldots,M)$ video sequences containing spatiotemporal words from a vocabulary of size $V(i = 1,2,\ldots,V)$. Each video d_j is represented as a sequence of N_j spatiotemporal words $w = (w_1,w_2,\ldots,w_{Nj})$. Then the process that generates each video d_j in the corpus using sLDA model is as follows:

(1) Draw the number of spatio-temporal words: $N_j \sim \text{Poisson}(\xi)$.
(2) Draw the mixing proportions of topics: $\theta \sim \text{Dir}(\alpha)$.
(3) For each of the N_j words w_n,
  (a) Draw topic assignment $z_n|\theta \sim \text{Mult}(\theta)$.
  (b) Draw a spatio-temporal $w_n|z_n,\beta_{1K} \sim \text{Mult}(\beta_{zn})$.
(4) Draw category label $c|z_{1:N}, \eta, \delta \sim GLM(\bar{z}, \eta, \delta)$, where $\bar{z} := (1/N)\sum_{n=1}^{N} z_n$ is the empirical topic frequencies.

The family of probability distributions corresponding to this generative process is depicted as a graphical model in Fig. 4. The distribution of the category label is a generalized linear model (GLM):

$$p(c|z_{1:N}, \eta, \delta) = h(c, \delta) \exp\left[\frac{\left(\eta^T \bar{z}\right) c - A\left(\eta^T \bar{z}\right)}{\delta}\right]. \quad (3)$$

In our case, we assume an over dispersed Poisson distribution for the category label where $h(c,\delta) = 1/c!$ and $A(\eta^T \bar{z}) = \exp\{\eta^T \bar{z}\}$.

Note that what distinguishes sLDA from the usual GLM is that the covariates are the unobserved empirical frequencies of the topics in the document. In the generative process, these latent variables are responsible for producing the words of the document, and thus, the response and the words are tied.

The following three computational problems need to be addressed in order to analyze the data with sLDA.

(1) *Posterior inference*, computing the conditional distribution of the latent variables at the document level given its words and the corpus-wide model parameters. The posterior is thus a conditional distribution of topic proportion θ and topic assignments $z_{1:N}$. Given a document and response, the posterior distribution of the latent variable is

$$p(\theta, z_{1:N}|w_{1:N}, c, \alpha, \beta_{1:K}, \eta, \delta^2) = \frac{p(\theta|\alpha)\left[\prod_{n=1}^{N} p(z_n|\theta)p(w_n|z_n, \beta_{1:K})\right] p(c|z_{1:N}, \eta, \delta^2)}{\int p(\theta|\alpha)d\theta \sum_{z_{1:N}}\left[\prod_{n=1}^{N} p(z_n|\theta)p(w_n|z_n, \beta_{1:K})\right] p(c|z_{1:N}, \eta, \delta^2)}. \quad (4)$$

The normalizing value is the marginal probability of the observed data (i.e., the document $w_{1:N}$ and the response c). As with LDA, it is not efficiently computable. Thus, we appeal to variational methods to approximate the posterior. We let π denote the set of model parameters, $\pi = \{\alpha,\beta_{1:K},\eta,\delta\}$ and $q(\theta,z_{1:N})$ denote a variational distribution of the latent variables. The lower bound is

$$\begin{aligned}
&\log p(w_{1:N}|\pi) \\
&\quad = \log \int_{\theta} \sum_{z_{1:N}} p(\theta, z_{1:N}, w_{1:N}|\pi) \\
&\quad = \log \int_{\theta} \sum_{z_{1:N}} p(\theta, z_{1:N}, w_{1:N}|\pi)\frac{q(\theta, z_{1:N})}{q(\theta, z_{1:N})} \\
&\quad \geq E[\log p(\theta, z_{1:N}, w_{1:N}|\pi)] - E[\log q(\theta, z_{1:N})],
\end{aligned} \quad (5)$$

where all expectations are taken with respect to $q(\theta,z_{1:N})$. This bound is called the evidence lower bound (ELBO), which we denote $L(\cdot)$. The first term is the expectation of the log of the joint probability of hidden and observed variables, and the second term is the entropy of the variational distribution $H(q) = -E[\log q(\theta,z_{1:N})]$. Variational inference proceeds by iteratively updating the variational parameters and finds a local optimum of ELBO. The resulting variational distribution $q(\theta,z_{1:N})$ is used as a proxy for the posterior.

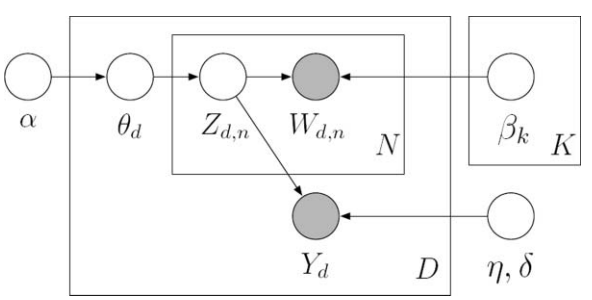


Fig. 4 sLDA graphical model. Nodes are random variables. Shaded ones are observed, and unshaded ones are unobserved. The plates indicate repetition.

(2) *Parameter estimation*, estimating the Dirichlet parameters α, the GLM coefficients η and the GLM dispersion parameter δ, and K topic multinomials $\beta_{1:K}$ from a data set of observed video-category pairs $\{w_{d,1:N}, c_d\}_{d=1}^{D}$. We fit these parameters with variational expectation maximization (EM), an approximate form of EM, where the expectation is taken with respect to a variational distribution. Note that each document is endowed with its own variational distribution. Expectations are taken with respect to that document-specific variational distribution $q(\theta,z_{1:N})$,

$$\begin{aligned}
&L(\alpha, \beta_{1:K}, \eta, \delta; D) \\
&\quad = \sum_{d=1}^{D} E_d[\log p(\theta_d, z_{d,1:N}, w_{d,1:N}, y_d)] + H(q_d).
\end{aligned} \quad (6)$$

In the expectation step (E-step), we estimate the approximate posterior distribution for each document-response pair using the variational inference algorithm described above. In the maximization step (M-step), we maximize the corpus-level ELBO with respect to the model parameters.

(3) *Prediction*, predicting a category label c from a newly observed video document $w_{1:N}$ and fixed values of the model parameters. This amounts to approximating the posterior expectation $E[w_{1:N},\alpha,\beta_{1:K},\eta,\delta]$. Given a new document $w_{1:N}$ and a fitted model $\{\alpha,\beta_{1:K},\eta,\delta\}$, we want to compute the expected response values,

$$\begin{aligned}
&E[c|w_{1:N}, \alpha, \beta_{1:K}, \eta, \delta] \\
&\quad = E[\mu(\eta^T \bar{Z})|w_{1:N}, \alpha, \beta_{1:K}].
\end{aligned} \quad (7)$$

Thus, given a new document, we first compute $q(\theta,z_{1:N})$, the variational posterior distribution of the latent variables θ and Z_n. We then estimate the response with

$$E[c|w_{1:N}, \alpha, \beta_{1:K}, \eta, \delta] \approx E_q[\mu(\eta^T \bar{Z})]. \quad (8)$$

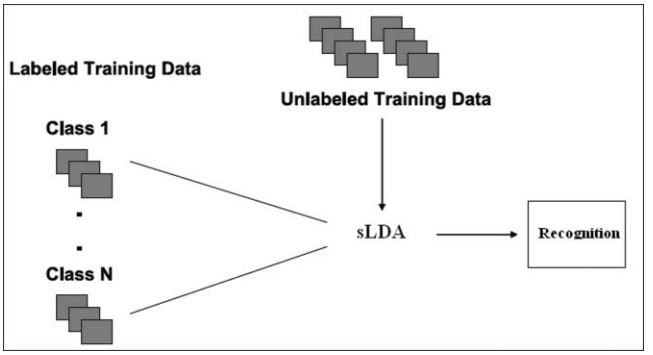


Fig. 5 Schematic overview of the hybrid learning algorithm using both labeled and unlabeled data.

3.3 *Hybrid Learning Using Labeled and Unlabeled Data*

Although the generalization performance of generative topic models is improved by training them discriminatively under the framework of the sLDA model, they can no longer make use of unlabeled data. In an attempt to gain the benefit of both generative and discriminative approaches, we extend the sLDA model to learn human action models from both labeled and unlabeled data. The proposed framework is shown in Fig. 5.

The generative process of corpus is the same as the sLDA. The differences lie in the parameter estimation procedure. When labeled data are available, we find topic structures that are more discriminative for the prediction of document categories. When unlabeled data are available, we treat the parameters of GLM as constant, use the model parameters estimated from labeled data, and only update the Dirichlet parameter α and the topic distributions $\beta_{1:K}$. The parameter-estimation step aims to maximize a generative likelihood of labeled and unlabeled data. This hybrid method allow semisupervised learning in the presence of strongly overlapping classes and reduce the risk of modeling structure in the unlabeled data that is irrelevant for the specific classification.

4 Experimental Results

We test our algorithm using three data sets: the KTH human motion data set,[34] the Weizmann human action data set,[12] and a ballet data set.[35] See Fig. 6 for sample frames from each data set. Our approaches are efficient. Most of the computation is spent on the features and code words. After the bag-of-words representation is obtained, learning the model usually takes less than 1 min, and inference on a new video only takes a few seconds in our unoptimized C code.

4.1 *Human Action Recognition Using the KTH Data Set*

The KTH human motion data set contains six types of human actions (walking, jogging, running, boxing, hand waving, and hand clapping) performed several times by 25 subjects in four different scenarios: outdoors, outdoors with scale variation, outdoors with different clothes, and indoors. It contains 598 short sequences.

We extract interest points and describe the corresponding spatiotemporal patches with the procedure described in Sec. 3.1. Examples of the detected features for sequences from the boxing and hand-waving categories are shown in Fig. 7. In order to build the codebook, we need to cluster the feature descriptors of all training video sequences. However, because the total number of features from all training examples is very large, we randomly select a subset of features to learn the codebook, in order to accommodate the requirements of memory. The data set has been partitioned into two randomly selected halves; one half forms the training set and

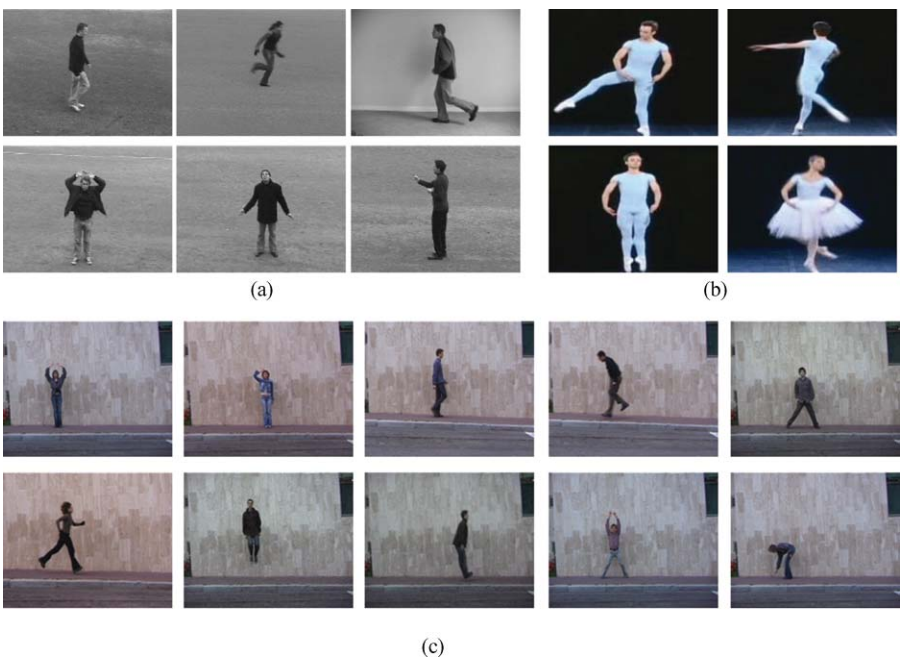


Fig. 6 Sample frames from our data sets. The action labels in each data set are as follows: (a) the KTH data set: walking, jogging, running, boxing, hand waving, hand clapping; (b) the Ballet data set: swinging, jumping, turning, hopping; (c) the Weizmann data set: bend, jack, jump, pjump, run, side, skip, walk, wave1, wave2.

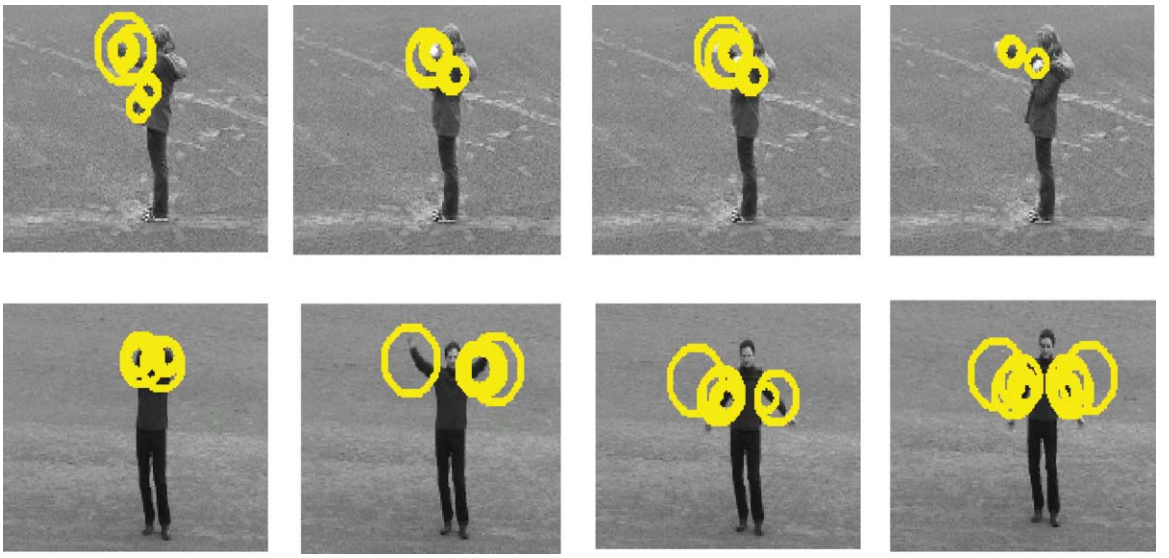


Fig. 7 Detected features on example frames of (first row) boxing and (second row) hand waving action category from the KTH dataset.

the other half forms the testing set. The training set consists of half of the videos from each action category. The remaining videos in the data set form the testing set for separate performance evaluation. In this way, videos in both the training and testing sets are randomly selected from the data set and there are videos from each category in both the training and testing sets. We repeat the classification experiments 40 times, and the average confusion matrix for sLDA on the KTH data set using 1000 spatiotemporal words is shown in Fig. 8(a). Each row in the confusion matrix corresponds to the ground-truth class, and each column corresponds to the assigned class label. We can see that the algorithm correctly classifies most actions. The confusion matrix shows the largest confusion between "jogging" and "running" actions. This is consistent with our intuition that similar actions are more easily confused with each other, such as those involving hand or leg motions.

We test the effect of the number of spatiotemporal words on recognition accuracy on both sLDA and unsupervised LDA models, as illustrated in Fig. 8(b), which shows some dependency of the recognition accuracy on the size of the codebook. Additionally, we can see sLDA performs better than unsupervised LDA because it employs discriminative training using labeled data.

To evaluate the performance of action recognition using both labeled and unlabeled data, the training set has been partitioned into two randomly selected halves; one half forms the labeled training set and the other half forms the unlabeled

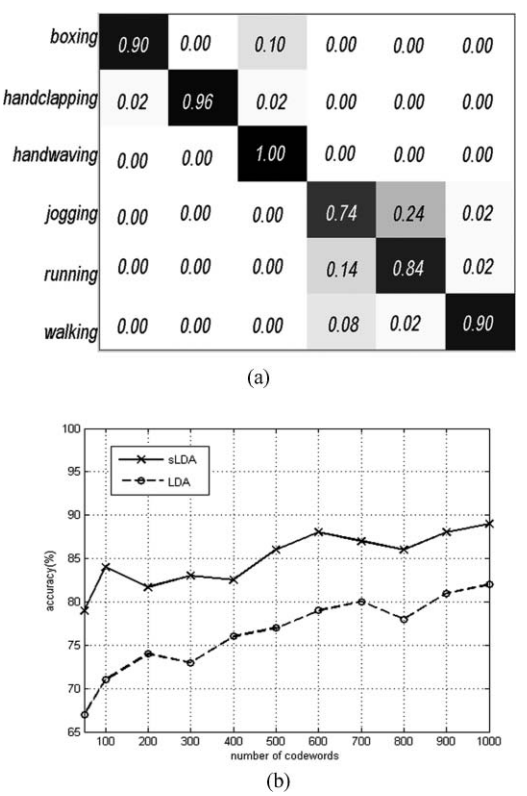


Fig. 8 (a) Confusion matrix for the KTH dataset using 1000 code words (overall accuracy = 89%); (b) classification accuracy versus codebook size.

Table 1 Comparison of different reported results on the KTH data set.

Methods	Recognition accuracy (%)
Our method (labeled data)	89.00
Our method (labeled and unlabeled data)	87.52
Niebles et al.[1]	83.33
Dollar et al.[15]	81.17
Wong et al.[30]	91.60
Nowozin et al.[36]	87.04
Wang and Mori[31]	91.20
Laptev et al.[34]	96.35
Ke et al.[16]	62.96

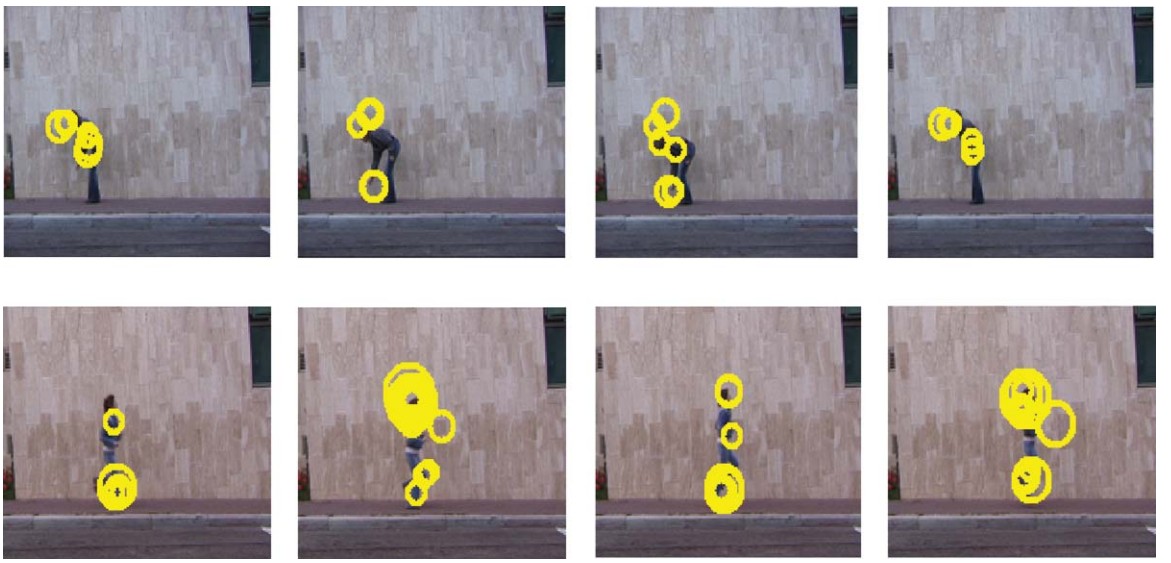


Fig. 9 Detected features on example frames of bend (first row) and skip (second row) action category from the Weizmann dataset.

training set. We compare our results to previous approaches on the same data set, as shown in Table 1. It should be noted that different methods listed in Table 1 have all sorts of variations in their experimental setups (e.g., different splits of training and testing data, whether some preprocessing is needed, with or without supervision, etc.). The results of our methods are comparable to other state-of-the-art approaches; although, we emphasize that this is not a precise comparison due to the variations in the experimental setup.

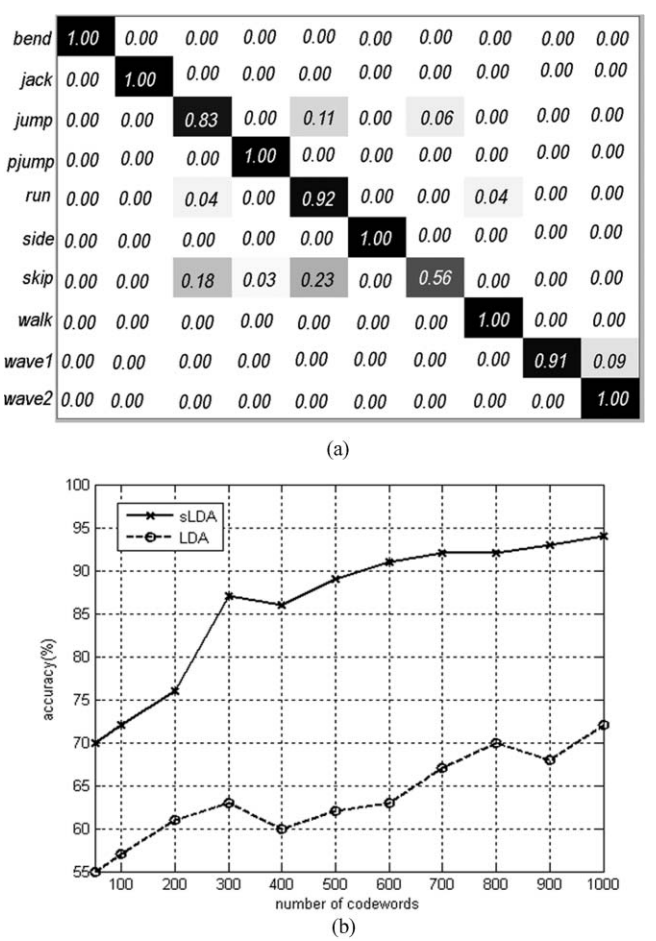


Fig. 10 (a) Confusion matrix for the Weizmann dataset usig 800 code words (overall accuracy = 92.20%); (b) classification accuracy versus codebook size.

4.2 *Human Action Recognition Using the Weizmann Data Set*

In our second experiment, we employ the Weizmann human action data set. It contains 10 action categories (bend, jack, jump, pjump, run, side, skip, walk, wave1, and wave2) performed by nine people, to provide a total of 90 videos.

We detect and describe spatiotemporal interest points using the procedure detailed in previous sections. Examples of the detected features for sequences from the bend and skip action category are shown in Fig. 9. The codebook is learned using all the feature descriptors obtained from all the training video sequences. We adopt a leave-one-out scheme to test the efficiency of our approach in recognition (i.e., for each run, we learn a model from the videos of eight subjects and test those of the remaining subject). The result is reported as the average of nine runs. The confusion matrix for a ten-class model is presented in Fig. 10(a) for the sLDA model learned using a codebook size of 800. The average performance with this codebook size is 92.20%. Note that the confusion matrix shows how our model is mostly confused by similar action classes, such as "skip" with "jump" and "run," or "wave1" with "wave2."

We test the effect of the number of spatiotemporal words on recognition accuracy on the sLDA and LDA models, as illustrated in Fig. 10(b). It shows some dependency of the recognition accuracy on the size of the codebook.

We also compare our results to previous methods in Table 2. We follow a similar way of learning actions using both labeled and unlabeled data in the training set. Again, we accept the fact that different methods have all sorts of variations in their experimental setups. For example, in Wang and Mori,[31] they need to track and stabilize the video sequences. In addition, their experimental results were reported using

Table 2 Comparison of different reported results on the Weizmann data set.

Methods	Recognition accuracy (%)
Our method (labeled data)	92.20
Our method (labeled and unlabeled data)	91.47
Niebles et al.[1]	90.00
Niebles and Fei-Fei[37]	72.80
Wang and Mori[31]	98.91

Table 3 Comparison of different reported results on the ballet data set.

Methods	Recognition accuracy (%)
Our method (labeled data)	96.25
Our method (labeled and unlabeled data)	93.72
Fathi and Mori[35]	51
Wang and Mori[31]	91.36

9 of the 10 action categories (excluding the most confusing "skip" action). Niebles et al.[1] obtain their best performance with pLSA using 1200 code words with an experimental setup similar to ours.

4.3 *Human Action Recognition Using the Ballet Data Set*

Finally, we test our algorithm on a ballet data set. We manually collect 86 video sequences from the dataset. Each video is labeled with one of the four action labels: swinging, jumping, turning, and hopping.

We extract interest points and describe the corresponding spatiotemporal patches with the procedure described in Sec. 3.1. Examples of the detected features for sequences in each category are shown in Fig. 11. The data set has been partitioned into two randomly selected halves; one half forms the training set and the other half forms the testing set. The training set consists of half of the videos from each action category. The rest videos in the data set forms the testing set for separate performance evaluation. In this way, videos in both the training and testing sets are randomly selected from the data set, and there are videos from each category in both the training and testing sets. The codebook is learned using all the feature descriptors obtained from all the training video sequences. We repeat the classification experiments 20 times, and the average confusion table for sLDA using 300 spatiotemporal words is shown in Fig. 12.

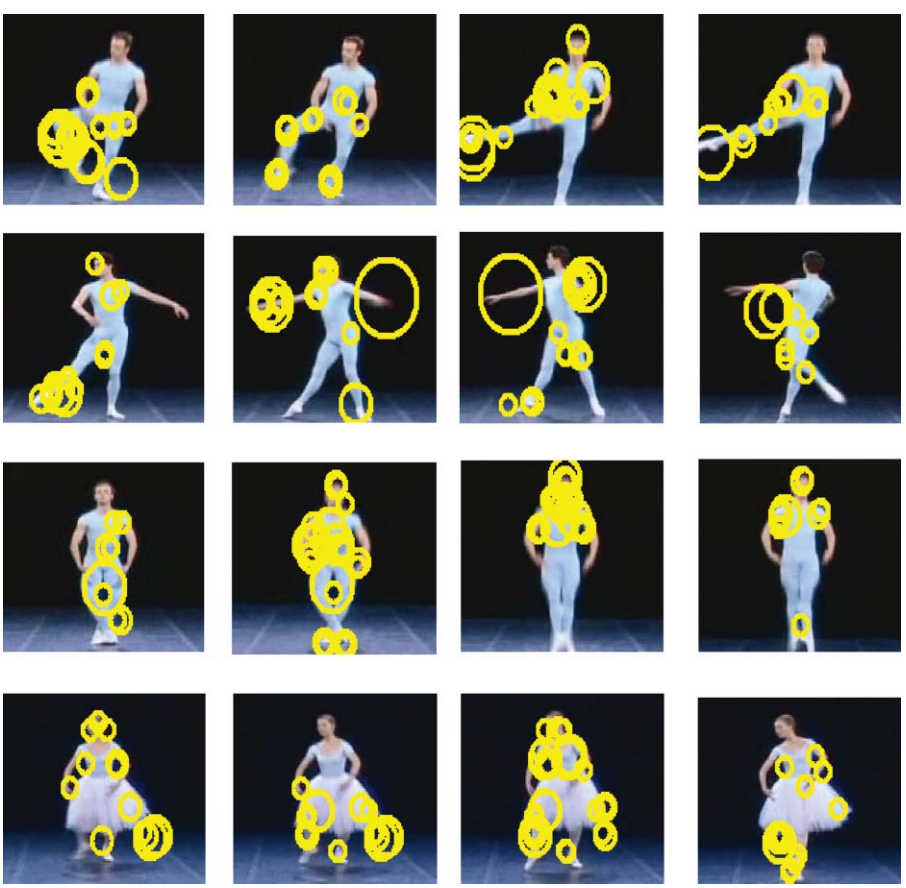


Fig. 11 Detected features on example frames of each action category from the Ballet data set.

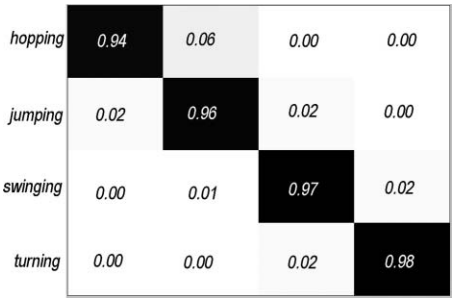


Fig. 12 Confusion matrix for the Ballet dataset using 300 code words (overall accuracy = 96.25%).

We compare our results to Fathi and Mori[35] and Wang and Mori[31] in Table 3. It should be noted that there are variations in the experimental setups.

5 Conclusions

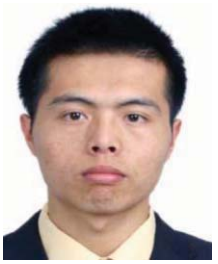
In this paper, we have presented a hybrid generative-discriminative learning approach for human action recognition from video sequences. Our model combines a bag-of-visual-words component with supervised topic discovery. We also extend the supervised topic model to learn human action with both labeled and unlabeled data. Experimental results on three challenging data sets show that the classification performance is on par with the current state-of-the-art results. The results are promising and reflect the promise of hybrid generative-discriminative learning approaches. In further research, we plan to learn multiple human action models in a single video sequence in realistic surveillance scenarios.

Acknowledgments

The authors thank the associate editor, Prof. Andrea Prati, and the anonymous reviewers for their valuable comments, which helped improve the clarity of the presentation of this paper.

References

1. J. C. Niebles, H. Wang, and L. Fei-Fei, "Unsupervised learning of human action categories using spatial-temporal words," *Int. J. Comput. Vis.* **79**, 299–318 (2008).
2. A. D. Holub, M. Welling, and P. Perona, "Hybrid generative-discriminative visual categorization," *Int. J. Comput. Vis.* **77**, 239–258 (2008).
3. L. Fei-Fei and P. Perona, "A Bayesian hierarchical model for learning natural scene categories," in *Proc. of IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 524–531 (2005).
4. J. Sivic, B. C. Russell, A. Efros, A. Zisserman, and W. T. Freeman, "Discovering objects and their location in images," in *Proc. of Tenth IEEE International Conference on Computer Vision*, pp. 370–377 (2005).
5. D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet allocation," *J. Mach. Learning Res.* **3**, 993–1022 (2003).
6. T. Hofmann, "Probabilistic latent semantic indexing," in *Proc. of 22nd Annual Int. Conf. on Research and Development in Information Retrieval*, pp. 50–57 (1999).
7. D. Ramanan and D. A. Forsyth, "Automatic annotation of everyday movements," in *Proc. of Advances in Neural Information Processing Systems*, pp. 1547–1553 (2004).
8. A. Yilmaz and M. Shah, "Recognizing human actions in videos acquired by uncalibrated moving cameras," in *Proc. of Tenth IEEE International Conference on Computer Vision*, pp. 150–157 (2005).
9. Y. Song, L. Goncalves, and P. Perona, "Unsupervised learning of human motion," *IEEE Trans. Pattern Anal. Mach. Intell.* **25**, 1–14 (2003).
10. C. Fanti, L. Zelnik-Manor, and P. Perona, "Hybrid models for human motion recognition," in *Proc. of IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 1166–1173 (2005).
11. A. F. Bobick and J. W. Davis, "The recognition of human movements using temporal templates," *IEEE Trans. Pattern Anal. Mach. Intell.*, **23**, 257–267 (2001).
12. M. Blank, L. Gorelick, E. Shechtman, M. Irani, and R. Basri, "Actions as space-time shapes," in *Proc. of Tenth IEEE International Conference on Computer Vision*, pp. 1395–1402 (2005).
13. A. A. Efros, A. C. Berg, G. Mori, and J. Malik, "Recognizing action at a distance," in *Proc. of Ninth IEEE International Conference on Computer Vision*, pp. 726–733 (2003).
14. I. Laptev, "On space-time interest points," *Int. J. Comput. Vis.* **64**, 107–123 (2005).
15. P. Dollar, V. Rabaud, G. Cottrell, and S. Belongie, "Behavior recognition via sparse spatio-temporal features," in *Proc. of 2nd Joint IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance*, pp. 65–72 (2005).
16. Y. Ke, R. Sukthankar, and M. Hebert, "Efficient visual event detection using volumetric features," in *Proc. of Tenth IEEE International Conference on Computer Vision*, pp. 166–173 (2005).
17. A. Oikonomopoulos, I. Patras, and M. Pantic, "Human action recognition with spatiotemporal salient points," *IEEE Trans. Syst. Man Cybern., Part B: Cybern.* **36**, 710–719 (2006).
18. T. Kadir and M. Brady, "Scale saliency: a novel approach to salient feature and scale selection," in *Proc. of Int. Conf. on Visual Information Engineering*, pp. 25–28 (2003).
19. A. Y. Ng and M. I. Jordan, "On discriminative vs. generative classifiers: a comparison of logistic regression and naïve bayes," in *Proc. of Advances in Neural Information Processing Systems*, pp. 841–848 (2002).
20. G. Bouchard and B. Triggs, "The trade-off between generative and discriminative classifier," in *Proc. of Computational Statistics the 16th Symp.*, pp. 721–728 (2004).
21. A. Holub and P. Perona, "A discriminative framework for modeling object classes," in *Proc. of IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 664–671 (2005).
22. A. Bar-Hillel, T. Hertz, and D. Weinshall, "Object class recognition by boosting a part-based model," in *Proc. of IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 702–709 (2005).
23. C. M. Bishop and J. Lasserre, "Generative or discriminative? Getting the best of both worlds," in *Bayesian Statistics*, J. M. Bernardo et al. (Eds), Oxford University Press, London, pp. 3–23 (2007).
24. J. H. Xue and D. M. Titterington, "Interpretation of hybrid generative/discriminative algorithms," *Neurocomputing* **72**, 1648–1655 (2009).
25. M. Fritz, B. Leibe, B. Caputo, and B. Schiele, "Integrating representative and discriminative models for object category detection," in *Proc. of Tenth IEEE International Conference on Computer Vision*, pp. 1363–1370 (2005).
26. B. C. Russell, A. A. Efros, J. Sivic, W. T. Freeman, and A. Zisserman, "Using multiple segmentations to discover objects and their extent in image collections," in *Proc. of IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 1605–1614 (2006).
27. R. Fergus, L. Fei-Fei, P. Perona, and A. Zisserman, "Learning object categories from Google's image search," in *Proc. of Tenth IEEE International Conference on Computer Vision*, pp. 1816–1823 (2005).
28. D. Liu and T. Chen, "A topic-motion model for unsupervised video object discovery," in *Proc. of IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 1–8 (2007).
29. A. Bissacco, M. H. Yang, and S. Soatto, "Detecting human via their pose," in *Proc. of Advances in Neural Information Processing Systems*, pp. 169–176 (2007).
30. S. F. Wong, T. K. Kim, and R. Cipolla, "Learning motion categories using both semantic and structure information," in *Proc. of IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 1–6 (2007).
31. Y. Wang and G. Mori, "Human action recognition by semilatent topic models," *IEEE Trans. Pattern Anal. Mach. Intell.* **31**, 1762–1774 (2009).
32. H. Wang, M. M. Ullah, A. Klaser, I. Laptev, and C. Schmid, "Evaluation of local spatio-temporal features for action recognition," in *Proc. of British Machine Vision Conf.*, pp. 127–138 (2009).
33. D. M. Blei and J. D. McAuliffe, "Supervised topic models," in *Proc. of Advances in Neural Information Processing Systems*, pp. 324–335 (2008).
34. I. Laptev, B. Caputo, C. Schultz, and T. Lindeberg, "Local velocity-adapted motion events for spatio-temporal recognition," *Comput. Vis. Image Understand.* **108**, 207–229 (2007).
35. A. Fathi and G. Mori, "Action recognition by learning mid-level motion features," in *Proc. of IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 1–8 (2008).
36. S. Nowozin, G. Bakir, and K. Tsuda, "Discriminative subsequence mining for action classification," in *Proc. of IEEE 11th International Conference on Computer Vision*, pp. 1–8 (2007).
37. J. C. Niebles and L. Fei-Fei, "A hierarchical model of shape and appearance for human action classification," in *Proc. of IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 1–8 (2007).



Hao Sun received his BSc and MSc in National University of Defense Technology, Changsha, China, in 2006 and 2008, respectively. He is currently pursuing a PhD in the university's School of Electrical Science and Engineering. His research interests include image analysis and understanding, pattern recognition and information fusion.

Cheng Wang received his BSc and PhD in communication and signal processing from the National University of Defense Technology, Changsha, China, in 1997 and 2002, respectively. He is currently a professor at Xiamen University. He is also the co-chair of the Working Group I/3 "Multi-Platform Multi-Sensor Inter-Calibration" in the International Society of Remote Sensing. His research interests include image analysis, information fusion, and mobile mapping data processing.

Boliang Wang is a professor in the Department of Computer Science at Xiamen University, Fujian, China. He is also a professor in the School of Electronic Science and Engineering, at National University of Defense Technology, Changsha, China. His area of expertise includes image processing, multisensor integration, and pattern recognition. He has published over 80 scientific papers and patents.