Dynamic scene understanding by improved sparse topical coding

Wei Fu, Jinqiao Wang*, Hanqing Lu, Songde Ma

National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China

Abstract

The explosive growth of cameras in public areas demands a technique which develops a fully automated surveillance and monitoring system. In this paper, we propose a novel unsupervised approach to automatically explore motion patterns occurring in dynamic scenes under an improved sparse topical coding (STC) framework. Given an input video, it is segmented into a sequence of clips without overlapping. Optical flow features are extracted from each pair of consecutive frames, and quantized into discrete visual flow words. Each video clip is interpreted as a document and visual flow words as words within the document. Then the improved STC is applied to explore latent patterns which represent the common motion distributions of the scene. Finally, each video clip is represented as a weighted summation of these patterns with only a few non-zero coefficients. The proposed approach is purely data-driven and scene independent, which make it suitable for very large range applications of scenarios, such as rule mining and abnormal event detection. Experimental results and comparisons on various public datasets demonstrate the promise of the proposed approach.

1. Introduction

Video camera networks become ubiquitous. It is reported that over 1.85 billion cameras have been installed in UK alone. With the proliferation of cameras in public areas, it is increasingly desirable to develop fully automated surveillance and monitoring systems, with which people will be free from strenuous and boring labor.

Kuettel et al. [1], summarized problems in scene understanding as three questions below: (1) What are the typical actions in the scene? (2) How do they relate to each other? (3) What are the rules governing the scene? In this paper, we focus on how to automatically learn semantic motion patterns for a dynamic scene. Discovering motion patterns will directly lead to a semantic scene model and further facilitate the task of scene analysis. However, it is still a challenging task in both computer vision and pattern recognition. Because motion patterns are not only related to low-level features but also used to reflect semantic information from videos, which is video understanding, a high-level task in computer vision.

Fig. 1 gives two examples of far-field traffic surveillance scenarios. Some typical activities, which we call motions patterns of scenes, occur regularly and periodically. For this type of complex or crowded scenarios, the performance of most existing technologies such as object detection, tracking and classification degenerates heavily. Instead, we explore the structured information of scene from a video sequence directly using low-level features. Then motion pattern discovering is formulated as a sparse topical coding (STC) problem. By learning a semantic dictionary, a dynamic scene can be reconstructed as the summation of bases with only a few non-zero coefficients.

In the past few years, there has been a major effort underway in the vision community to pursue a sparse representation for images and videos. Significant work has shown the power of a sparse representation for a vision task, such as image classification [2] and annotation [3], action recognition [4], abnormal detection [5,6] and so on. However, representations learned by their methods seem to lack significant semantics, which could not be directly used in semantic scene modeling.

To address the problems above, in this paper, an improved sparse topical coding framework is proposed to learn a motion pattern dictionary. Then a dynamic scene can be represented as a weighted summation of these patterns with only a few non-zero coefficients. As we will show in our experiments, the bases learned by our approach have explicit semantics. Additionally, this mid-level sparse representation for the video, aiming at bridging the semantic gap of low-level features and high-level concepts, can be further applied to scene analysis such as rule mining, abnormal event detection and so on.

The main contributions of this paper can be summarized as follows:

1. In our method, the problem of dynamic scene understanding is addressed under an improved sparse topical coding framework, which can automatically explore motion patterns occurring in dynamic scenes.

2. Each video clip is treated as a mixture of topics (motion patterns). By directly imposing a sparse bias on the distribution

* Corresponding author. Tel.: +86 10 62527720.
E-mail addresses: wfu@nlpr.ia.ac.cn (W. Fu), jqwang@nlpr.ia.ac.cn (J. Wang), lhq@nlpr.ia.ac.cn (H. Lu), masd@most.cn (S. Ma).
of topics, a dynamic scene can be sparsely reconstructed with a dictionary.
3. Compared with other methods (LDA [7] and HDP [1]), our improved STC achieves better performance in both scene pattern discovering and abnormal event detection.

2. Related work

Dynamic scene understanding has become an active area especially in video surveillance. Various attempts have been made to address the three problems mentioned above. In this section, we will review state-of-the-arts from two categories: object based approaches [8,9] and non-object based approaches [10–12,7,1].

The first kind of approaches treat scene analysis as a traditional object detection and tracking problem. For example, Zhang et al. [8] trained classifiers under a co-training framework to classify object into different classes, and they cluster each class of trajectories according to their spatial distributions. Li et al. [9] detected moving objects as video events. Feature similar video events are grouped into atomic events, which were further clustered into behaviors using GMM and EM algorithms. Both approaches above suffer difficulties in crowded scenes, where the use of tools typically used in surveillance.

To handle these difficulties, some researchers [10,11] directly use the low-level motion as appearance features. Instead of object detection, tracking and classification, beginning with low-level features, they usually employ a Dynamic Bayes network to explore the structure and temporal information of videos. Li et al. [11] employed a statistical model of optical flow features to obtain a representation for the motion patterns. Another two most related to our works, Wang et al. [12] and Kuettel et al. [1], characterized typical activities by hierarchical Bayesian models, such as latent Dirichlet allocation (LDA) and hierarchical Dirichlet process (HDP). Although a similar topic model is employed in our scene understanding problem. Before introducing our proposed approach, we firstly give a brief overview of the topic model and in particularly the sparse topical coding model.

Researchers [1,12] have put them forward to address the problems in computer vision community. In order to learn a representation that captures the latent semantics of a large collection of data, Zhu and Xing [16] presented a non-probabilistic formulation of topic model—sparse topical coding.

Given a collection of D documents \( D = \{ w_1, w_2, \ldots, w_D \} \) and a vocabulary with N total words denoted as \( V = \{ 1, 2, \ldots, N \} \), we represent each document \( w \) as a vector \( w = (\omega_1, \omega_2, \ldots, \omega_N) \), where \( I \) is the set of word index and \( \omega_n \) denotes the number of occurrence of word \( n \) in this document. Let \( \beta \in \mathbb{R}^{K \times N} \) be a dictionary with \( K \) bases, and each row \( \beta_i \) is a distribution over the vocabulary \( V \). Then the goal of a topic model is to explore the abstract “topics” in the collection of data.

The STC model introduces two variables: \( \theta_d \in \mathbb{R}^K \) denoting the code of document \( d \) and \( s_{dn} \in \mathbb{R}^K \) standing for the code of word \( n \). STC assumes that (1) the word codes \( s_{dn} \) are conditionally independent given its document code \( \theta_d \), and (2) the observed word counts are independent given their latent representations \( s_{dn} \). With the conditionally independent assumptions, a generative procedure can be summarized as follows:

1. sample a dictionary \( \beta \) from a uniform prior distribution \( p(\beta) \).
2. for each document \( d \in \{ 1, \ldots, D \} \)
   (a) sample the document code \( \theta_d \) from a Laplace prior \( p(\theta_d) \).
   (b) for each observed word \( n \in I_d \)
      i. sample the word code \( s_{dn} \) from a conditional distribution \( p(s_{dn} | \theta_d) \).
      ii. sample the observed word count \( \omega_n \) from a conditional distribution \( p(o_n | s_{dn}, \beta_n) \).

In order to achieve sparse representations for document codes \( \theta_d \) and word codes \( s_{dn} \), STC chooses the Laplace prior \( p(\theta_d) \propto \exp(-\gamma |\theta_d|_1) \) and the supergaussian \( p(s_{dn} | \theta_d) \propto \exp(-\gamma |s_{dn} - \theta_d|_2^2 + \rho |s_{dn}|_1) \). Then STC is to minimize the following objective function:

\[
\min_{\theta, s, \beta} \sum_{d=1}^{D} \log p(o_n | s_{dn}, \beta) + \sum_{d=1}^{D} |\theta_d|_1 + \sum_{d, n \in I_d} (\gamma |s_{dn} - \theta_d|_2^2 + \rho |s_{dn}|_1).
\]

3. Theoretical background

Inspired by the research and ideas from the field of textual document analysis, common themes are extracted to represent motion patterns in our scene understanding problem. Before introducing our proposed approach, we firstly give a brief overview of the topic model and in particularly the sparse topical coding model.

Topic models (such as PLSA [13], LDA [14] and HDP [15]) were first developed in nature language processing to discover latent topics that occur in a large collection of documents. Recently, some
\[
\text{s.t. } \theta_\ell \geq 0 \ \forall \ell; \quad s_{dn} \geq 0 \ \forall d, \ n \in I_d; \quad \beta_k \in P \ \forall k; \quad (1)
\]

where \((\lambda, \gamma, \rho)\) are non-negative hyper-parameters set by users. The formulation of STC directly controls the sparsity of inferred representations by imposing sparsity-inducing regularizers on \(\theta\) and \(s\). Fig. 2 presents a graphical representation of the STC model.

### 4. Video representation by a topic model

Given an input video under a surveillance scenario, we first segment the whole video into a sequence of clips (documents) without overlapping. Optical flow features are extracted from each pair of consecutive frames, and quantized into discrete visual words. Then the video is represented by a word-document hierarchical topic model through a generative process.

Flow words. As done in [1, 12], we compute local motion as our low-level features. Given an input video, optical flow features are extracted for each pair of consecutive frames using the method proposed in [17]. A threshold is used to remove noise and only the distinctive motion pixels are preserved to be further processed. In order to generate the vocabulary, we sample optical flow vectors \(x = (x, y, u, v)\), whose position \((x, y)\) arranged on a grid with a spacing of 10 pixels. Then the sampled flow vectors are quantized into 8 directions according to their displacements \((u, v)\). Finally a fixed vocabulary is formed, in which each word contains two aspects of contents: position information and motion direction information. Similar as in a topic model, let \(V = \{1, 2, \ldots, N\}\) denote the vocabulary with \(N\) total flow words.

Motion patterns. In this paper, we define a motion pattern as a spatial distribution of flow words, which have high co-occurrence frequencies in the same video clip. Motion patterns, denoted as \(\beta\), are corresponding to the latent topics in a topic model. Each row of \(\beta\) is a topic basis, which is a distribution over the vocabulary \(V\), i.e., \(\beta_k \in P\), where \(P\) is a \((N-1)\)-simplex.

Mixture of patterns. Since the whole input video is divided into a sequence of clips without overlapping, each clip is treated as a document and the whole video corresponds to the corpus in the topic model. Flow words within a video clip are accumulated over its frames. And then a clip of video is represented as a vector \(w = (\omega_1, \ldots, \omega_l)^T\), where \(l\) is the set of word indexes and \(\omega_n\) denotes the number of occurrence of word \(n\) in this clip.

Obviously, each document (video clip) is a mixture of topics (motion patterns). Here, we denote \(\theta\) for the code of the set of clips. In our approach, we pursue a sparse representation for each \(\theta_k\), that means, a video clip can be sufficiently interpreted by only a few motion patterns which most impossibly occur during this time.

\[f(\Theta; \beta) = \sum_{d,n \in I_d} (s_{dn} \beta_n) + \lambda_1 \sum_{n \in I_d} \| \beta_n \|_1 + \lambda_2 \sum_{d} \| \theta_d \|_1 + \sum_{d,n \in I_d} \langle s_{dn} \theta_d, \beta_n \rangle^2 + \rho \| s_{dn} \|_1 \]

\[\text{s.t. } \theta_\ell \geq 0 \ \forall \ell; \quad s_{dn} \geq 0 \ \forall d, \ n \in I_d; \quad \beta_k \in P \ \forall k; \quad (4)\]

where \(\Theta = \{\theta_\ell, s_{dn}\}_{d=1}^D\) denotes the codes for a collection of documents \(\{w_d\}_{d=1}^D\), and \(f(\Theta; \beta) = -\log p(\omega_n | s_{dn}, \beta)\) is a cost function, aiming at minimizing the unnormalized KL-divergence between observed word counts \(\omega_n\) and their reconstructions \(s_{dn} \beta_n\). And \(\lambda_1, \lambda_2, \gamma, \rho\) are non-negative hyper-parameters set by users. This is what we call improved STC.

### 5. Learning motion patterns

We try to learn a sparse representation that captures the latent semantics of a large collection of video data. Sparse topical coding [16] could learn semantic topical bases with a two layer model. In this section, we will detail how to formulate the problem of learning motion patterns under the STC framework. In addition, an improvement for the standard STC is proposed to fit our model. Model learning for our improved STC is presented accordingly.

#### 5.1. Problem formulation

After representing a video by a topic model, according to the generative procedure described in Section 3, a joint probability distribution can be defined as follows:

\[p(\theta, s, w, \beta) = p(\beta)p(\theta|\beta)\prod_{d \in I_d} p(s_d|\theta)p(\omega_n|s_d, \beta). \quad (2)\]

It is reasonable that a word occurs with an average rate and independently of the time since the last one. According to the probability and statistics theory, we assume the discrete word counts obey a Poisson distribution with \(s_d \beta_n\) as the mean parameter, i.e.,

\[p(\omega_n|s_d, \beta) = \frac{(s_d \beta_n)^{\omega_n} \exp(-s_d \beta_n)}{\omega_n!}. \quad (3)\]

In the standard STC model, a uniform distribution is assumed to \(\beta\) (see Fig. 2). However, in most surveillance scenes, motion patterns often occur with different frequencies. That means, the importance of each motion pattern is different. To capture this characteristic of dynamic scenes, we argue that words (flow words) in all documents (video clips) should concentrate on some dominant topics (motion patterns). To this end, we add an \(\ell_1\)-norm constraint term for \(\beta_n\) to the standard STC formulation. We will show that this constraint term is also helpful to determine an optimal number of topics in Section 7. Finally, the objective function of the convex problem can be rewritten as

\[f(\Theta; \beta) = \sum_{d,n \in I_d} (s_{dn} \beta_n) + \lambda_1 \sum_{n \in I_d} \| \beta_n \|_1 + \lambda_2 \sum_{d} \| \theta_d \|_1 + \sum_{d,n \in I_d} \langle s_{dn} \theta_d, \beta_n \rangle^2 + \rho \| s_{dn} \|_1 \]

\[\text{s.t. } \theta_\ell \geq 0 \ \forall \ell; \quad s_{dn} \geq 0 \ \forall d, \ n \in I_d; \quad \beta_k \in P \ \forall k; \quad (4)\]
5.2. Model learning by improved STC

The formulation (4) leads to a semi-convex optimization problem, that is, \( f(\Theta, \Phi) \) is convex over either \( \Theta \) or \( \Phi \) when the other is fixed. Similar as in [16], we utilize an effective method to solve the problem (1), which alternately performs hierarchical sparse coding and dictionary learning procedure to solve this problem. Different from [16], we need consider a sparsity constraint when learning the dictionary \( \Phi \). To be specific, the whole optimization procedure is carried as below:

Hierarchical sparse coding: When the dictionary \( \Phi \) is fixed, this step finds optimal \( s \) and \( \theta \) by alternatively solving two problems below: For \( s \), when \( \theta \) is fixed, we solve each \( s_n \) by solving

\[
\min_{s_n} \ell(s_n, \Phi) + \gamma ||s_n - \theta||_2^2 + \rho \sum_k s_{nk}, \quad \text{s.t. } s_n \geq 0. \tag{5}
\]

For \( \theta \), when \( s \) is fixed, we solve the problem

\[
\min_{\theta} \lambda ||\theta||_1 + \gamma \sum_n ||s_n - \theta||_2^2, \quad \text{s.t. } \theta \geq 0. \tag{6}
\]

Under the non-negative constraints to \( \theta \) and \( s_n \), both the two problems have closed form solutions, which can be effectively computed. Refer to [16] for details.

Sparse dictionary learning: Once the representations \((\theta, s)\) of all the documents are inferred, the dictionary \( \Phi \) is learned by minimizing the following objective function:

\[
\min_{\Phi} \ell(s_n, \Phi) + \lambda ||\Phi||_1 \tag{7}
\]

In this problem, the first term, a log-Poisson loss function, is convex but non-quadratic. To minimize it with an \( \ell_1 \) constraint, Bregman iteration algorithm [18] is utilized here. We introduce a slack variable \( b_n = s_n^T \Phi \) and then for each column of \( \Phi \) the formulation becomes:

\[
\min_{b_n, \Phi_n} b_n - w_n \log b_n + \lambda ||\Phi_n||_1 \quad \text{s.t. } b_n = s_n^T \Phi_n, n = 1, \ldots, N. \tag{8}
\]

Fig. 4. The first 9 most important patterns discovered by our improved STC, shown in order of decreasing importance.
Further, we have a Bregman iteration formulation as follows:

\[
(b_{a}^{k+1}, \beta_a^{k+1}) = \arg \min_{b_a, \beta_a} -w_a \log b_a + \lambda_1 \| \beta_a \|_1 + \varepsilon \|s_a^T \beta_a - b_a + p_k\|_2^2,
\]

where we alternately optimize \( \beta_a \) and \( b_a \) as below:

\[
b_{a}^{k+1} = \arg \min_{b_a} -w_a \log b_a + \varepsilon \|s_a^T \beta_a - b_a + p_k\|_2^2,
\]

\[
\beta_a^{k+1} = \beta_a^{k} + (s_a^T b_{a}^{k+1} - b_{a}^{k+1}).
\]

The first optimization Problem (11) has a closed form solution, and the second one can be solved as a lasso problem.

Now, we summarize the whole procedure of model learning as Algorithm 1.

Algorithm 1. The procedure of model learning.

**Input:** Video clips \( \{w_d\} \), the number of topics \( K \), the hyperparameters \( (\lambda_1, \lambda_2, \gamma, \rho) \)

**Output:** the dictionary \( \beta \), document codes \( \theta \), word codes \( s \)

**Repeat**

1. **Hierarchical sparse coding:**
   for \( d = 1 \) to \( D \) do
     for each word \( n \) do
       Update word code \( s_n \) by solving Problem (5);
       Update document code \( \theta_d \) by solving Problem (6);
   end for
   end for

2. **Dictionary learning:**
   Update the dictionary \( \beta \) by solving Problem (7);
   Until convergence;

**6. Scene understanding with motion patterns**

In this section, we will discuss the task of scene understanding with the motion patterns learned before.

![Fig. 5. The first 9 most important patterns learned by HDP [1], shown in order of decreasing importance.](image)
Fig. 6. The first 9 most important patterns learned by the standard STC, shown in order of decreasing importance.

Fig. 7. Sparsity comparisons of document codes learned by improved STC, standard STC and HDP.
6.1. Scene rule mining

Generally speaking, a surveillance scene usually has its own structured layout, such as lanes, zebra crossings and interactions. And motion patterns are related to and dependent on each other both temporally and spatially. Both spatial dependency and temporal dependency make a dynamic scene follow a rule. By analyzing dependencies of the motion patterns, these information can be further inferred.

**Spatial dependency.** As described before, a motion pattern is a spatial distribution over flow words, that means, flow words within the same pattern have a high co-occurrence frequency, which reflects spatial dependencies among the visual words. In addition, there exist dominant motion patterns within each video clip, with which the video clip can be sparsely reconstructed. So reconstruction coefficients also reflect the spatial dependencies among dominant patterns.

**Temporal dependency.** As the whole video is finally represented by a sequence of document codes in chronological order, motion patterns in different video clips may be dependent on each other in time. In our setting scenarios, temporal dependencies usually reflect in the causal relationships between different patterns and the recurrences of a certain pattern with periodicity.

In practice, a simplification is carried to explore the dependencies of motion patterns. For each video clip, the topic mixture weights are threshold so that each clip has no more than one dominant motion pattern. Therefore the whole video is denoted as a sequence of dominant patterns. Then Markov Models are utilized to learn the transitions of these patterns, which represent the scene rules.

6.2. Abnormal event detection

Abnormal event detection is a primary goal of research in video surveillance. The abnormal events can be identified as irregular events, which may be either spatial irregularly distributed or with abnormal temporal transitions. In former methods such as [5,6], an event is usually represented as features within an image patch or
spatial-temporal volume of frames, such as crowd escaping, cat falling down stairs, jet flipping and so on. In contrast, events in our approach are related directly to the patterns of motion over a period of time. Patterns deviating from the normal types are treated as abnormal events. With the motion patterns learned before, a sparse reconstruction cost (SRC) based approach is proposed.

Assuming a dictionary $\mathbf{b}$, each dynamic scene can be sparsely reconstructed by its bases. Then for a video clip with document code $\mathbf{h}$ and word codes $\mathbf{s}$, a sparse reconstructed cost (SRC) is defined as follows:

$$
\text{f}_{\text{SRC}} = \sum_{n \in I} \lambda \|\theta\|_1 + \sum_{n \in I} (\gamma \|s_n - \theta\|_2^2 + \rho \|s_n\|_1). 
$$

(14)

Obviously, a clip with a high SRC tends to be with abnormalities. The procedure of abnormal event detection can be summarized as Algorithm 2.

**Algorithm 2.** Abnormal event detection.

**Input:** A video clip represented as $\mathbf{w} = (w_1, \ldots, w_I)^T$, the dictionary learned before $\mathbf{b}$, an abnormal threshold $Th$

**Output:** the document codes $\mathbf{h}$, word codes $\mathbf{s}$

**Repeat**

for each word $n \in I$ do

- Update word code $s_n$ by solving Problem. (5);
- Update document code $h$ by solving Problem. (6);

Until convergence;

Compute $f_{\text{SRC}}$ according to Eq. (14);

An abnormal event is detected if $f_{\text{SRC}} > Th$;

7. Experiments

7.1. Datasets and settings

We evaluate the performance of our proposed approach on various complex and crowded public datasets. The first QMUL Junction video [9] is captured at 25 frames per-second, from a road junction. The video is approximate 60 min long (89999 frames), with a frame size of 360 x 288 pixels. The second ETH Traffic video [1], with a length of near 1 h (89 351 frame), is recorded under a dynamic scene with many agents, at 25 frames per-second and with a frame size of 960 x 540 pixels.

We first divide each video into 3 s length clips. Then optical flow vectors of each clip are quantized to discrete words in the vocabulary as described in Section 4. Then improved STC is utilized to learn semantic bases. For the first dataset, the hyper-parameters are set to be $(\lambda_1 = 500, \lambda_2 = 0.5, \gamma = 0.2, \rho = 0.2)$ while for the second dataset, the hyper-parameters are set to be $(\lambda_1 = 500, \lambda_2 = 0.35, \gamma = 0.1, \rho = 0.1)$.

In practice, without a prior knowledge for the topic number, usually a larger value (e.g., 20–30 in our experiments) is set to $K$. With a sparsity-inducing term on the dictionary, flow words are enforced to concentrate on dominant patterns with semantic concepts. Then these most informative motion patterns are selected according to their occurrence frequencies.

7.2. Experimental analysis

Semantic patterns discovering. For the QMUL Junction video, we set $K=20$ and we comprise our results with the baseline HDP [1]...
method. For clarity, we visualize the top 9 most important motion patterns learned by our improved STC, the HDP approach and the standard STC in Figs. 4–6 respectively.

Compared with the HDP approach, more patterns with explicit semantics (17 out of 20) are discovered by the improved STC. For example, the downward traffic lane like id1, the upward traffic lane like id3 and id4, the leftward traffic like id16, and so on. Besides, some patterns are comprised by other simpler patterns, such as id2 which is composed by a downward pattern and an upward pattern. Another example is id17, composed by a leftward traffic and turning right traffic. All these motion patterns compose to a dictionary for reconstructing video clips. Meanwhile, only 8 patterns have semantics in 21 total patterns learned by HDP.

Compared with the standard STC, the improved STC can explain more content of the video. From the results, we can find that, using the 9 dominant patterns, the standard STC can interpret 55% content of the video while the improved STC can interpret 82% content of the video. In fact, to interpret 80% content of a video, the standard STC needs at least 13 patterns. This means, the sparse dictionary makes the words concentrate on some topics, which discriminates the explored motion patterns.

Sparsity of document codes. Here, we check the sparsity of document codes $|\theta_{id}|$. Fig. 7 gives three exemplars to validate the sparsity of the mid-level representations for videos. Again, we compare our approach with the standard STC and the HDP method [1].

While the baseline HDP method assumed a sparse prior (i.e., Dirichlet) over the topics, our approach directly imposes a bias over the posterior probabilistic distributions, which has a more explicit intuition on the inferred representations. As Fig. 7 shows, on the premise of more semantic bases, our improved STC achieves the best performance in terms of sparsity of document codes.

This can be easily interpreted with the differences between the two dictionaries learned by improved STC and HDP. Our dictionary learned by improved STC contains not only simple bases but also compositional bases while bases of the dictionary learned by HDP are all simpler ones. With the compositional patterns, fewer bases are utilized in our method to reconstruct a complex scene.

Sparsity of the dictionary. Although the number of topics $K$ is determined by users initially, we find that the sparsity constraint for the dictionary did help us to cut off some non-meaningful bases. Due to the $\ell_1$ constraint term added to the dictionary, all visual words concentrate on certain topics with semantic concepts. We give an experimental result about the relationship between the coefficient of $|\beta_{nm}|$ and meaningful topic number. Here, we initialize $K=30$, as shown in Fig. 8, words concentrate on fewer semantic topics as $K$ increases.

Scene rule mining. For the ETH Traffic video, we set the topic number $K=30$ and give the first 9 most important motions according to their occurrence frequencies, as shown in Fig. 9. Notice that the horizontal flow (id8) is sometimes interrupted by the down-driving tram (id28), our approach also finds patterns like id3 and id6.

Fig. 10 shows the state transition matrix for the 9 patterns. From Fig. 10, we can infer the rule of pattern transitions. For example, id3–id6–id30 is a normal transition while id1–id28 is abnormal, which are consistent with the reality.

Abnormal event detection. In order to evaluate the performance of abnormal event detection with our improved STC model, we segment the QMUL Junction dataset into two parts equally. And the former part (including 600 clips) is trained to learn a dictionary while the latter one (including 599 clips) is used for testing. To give a quantitative evaluation, we manually label abnormal events for the testing data, including jaywalking, illegal U-turns, car cutting off and so on. In our experiments, 42 clips with abnormalities are totally found, and the ROC curves for the abnormal event detection are shown in Fig. 11. Compared with the standard STC, the improved STC obtains a lower false positive rate. This comparison result indicates that motion patterns learned by the improved STC are more distinctive for abnormality detection.

Five typical abnormal clips detected by our approach are illustrated in Fig. 12. As shown in Fig. 12a, two pedestrians jaywalk towards different directions. A motorbike turns suddenly off the horizontal route in Fig. 12b. A pedestrian heading towards an oncoming van is highlighted in Fig. 12c. An illegal u-turn and a jaywalking are detected in Fig. 12d and Fig. 12e, respectively.

8. Conclusions

In this paper, a novel unsupervised approach is proposed to learn semantic motion patterns for a dynamic scene. By representing a video as a topic model, an improved sparse topical coding framework is used to discover the semantic topical bases, with which each video clip can be sparsely reconstructed. Experimental results have shown the advantages of our approach. This work indicates that the sparse representations for videos are promising for scene analysis applications, such as scene rule mining and abnormal event detection.

Acknowledgment

This work was supported by 973 Program (2010CB327905) and National Natural Science Foundation of China (61070104, 60905008 and 61273034).

References

Wei Fu received his B.S. degree from University of Science and Technology of China (USTC), Hefei, China, in 2009. He is currently a Ph.D. candidate at National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing, China. His current research interests include pattern recognition and machine learning, image and video processing, and intelligent video surveillance.

Jinqiao Wang received the B.E. degree in 2001 from Hebei University of Technology, China, and the M.S. degree in 2004 from Tianjin University, China. He received the Ph.D. degree in pattern recognition and intelligence systems from the National Laboratory of Pattern Recognition, Chinese Academy of Sciences, in 2008. He is currently an associate professor with Chinese Academy of Sciences. His research interests include pattern recognition and machine learning, image and video processing, mobile multimedia, and intelligent video surveillance.

Hanqing Lu received his B.E. degree in 1982 and his M.E. degree in 1985 from Harbin Institute of Technology, and Ph.D. degree in 1992 from Huazhong University of Sciences and Technology. Currently, he is a deputy director of National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences. His research interests include image and video analysis, medical image processing, object recognition, etc. He has published more than 200 papers in these fields.

Songde Ma received his B.S. in Automatic Control from the Tsinghua University in 1968, Ph.D. degree in University of Paris in 1983 and “Doctorat d’Etat es Science” in France in 1986 in image processing and computer vision. He was an invited researcher in Computer Vision Laboratory in the University of Maryland in USA in 1983. From 1984 to 1986, he was a researcher in Robot Vision Laboratory in INRIA, France. Prof. Ma was a member of the Expert Committee of the National High Technology Program and the chief scientist of the Project “Image, Voice and Natural Language Understanding” of the National Fundamental Research Program. His research interests include computer vision, image understanding and searching, robotics and computer graphics, etc.