Copyright ⓒ2016 KSII

Crowd Activity Classification Using Category Constrained Correlated Topic Model

Xianping Huang[1], Wanliang Wang[1] , Guojiang Shen[1], Xiaoqing Feng[2] and Xiangjie Kong[3]
[1] College of Computer Science and Technology, Zhejiang University of Technology
Hangzhou, 310023 - China
[e-mail: {hxp,wwl, gjshen1975}@zjut.edu.cn]
[2] College of Information, Zhejiang University of Finance &Economics,
Hangzhou, 310018– China
[fenglinda@zufe.edu.cn]
[3] School of Software, Dalian University of Technology
Dalian, 116620 - China
[e-mail: xjkong@ieee.org]
*Corresponding author: Xianping Huang

Received April 10, 2015; revised December 17, 2015; accepted September 22, 2016; published November 30, 2016

Abstract

Automatic analysis and understanding of human activities is a challenging task in computer vision, especially for the surveillance scenarios which typically contains crowds, complex motions and occlusions. To address these issues, a Bag-of-words representation of videos is developed by leveraging information including crowd positions, motion directions and velocities. We infer the crowd activity in a motion field using Category Constrained Correlated Topic Model (CC-CTM) with latent topics. We represent each video by a mixture of learned motion patterns, and predict the associated activity by training a SVM classifier. The experiment dataset we constructed are from Crowd_PETS09 bench dataset and UCF_Crowds dataset, including 2000 documents. Experimental results demonstrate that accuracy reaches 90%, and the proposed approach outperforms the state-of-the-arts by a large margin.

Keywords: Human activity, Crowd surveillance, Bag-of-visual-words, CC-CTM

This work is partially supported by the Natural Science Fund of China (NSFC-61202197), the Science and Technique program of Zhejiang Province China (2015C31059) , the Science and Technique program of Zhejiang Provincial Department of transportation, China(2014T08) and the Public Welfare Technology and Industry Project of Zhejiang Provincial Science Technology Department (No.2016C31081). The authors would like to thank the researchers being with the college of computer Science and technology, Zhejiang University of Technology, China for his/her help.

http://dx.doi.org/10.3837/tiis.2016.11.018 **ISSN : 1976-7277**

1. Introduction

In recent years, more and more surveillance systems are equipped in public, such as airports, train stations and large squares. Automatic analysis of human activities in videos has attracted increasing attention for the purpose of public security, crowd management or public area aided design. Most existing methods can only handle videos which contain up to a number of people, and reconizing activities in videos containing hundreds or even thousands of people (like scenes in **Fig. 1**) is still a challenge task.

a. Structured crowd motion **b.** Unstrucured crowd motion

Fig. 1. Examples of crowd scenes. Note these scenes typically contains more than 100 people.

In order to process videos containing many people, some intrinsic properties of crowd scenes are explored. For example, people in crowd may move with homogeneous dynamics (**Fig. 1 (a)**), which can be modeled to characterize the crowd flow called as structured crowd motion [1]. Nevertheless, many scenes, known as unstructured crowd motion [2] contain partially or completely random motion of people, which complicates the analysis of the crowd activity. These scenes usually involve difficulties like serious occlusion, similar appearance between different activity classes and small-sized objects, which renders tracking and event detection approaches [4,5,6,7,8] inapplicable.

Modeling crowd motion [9] provides important priors for analyzing human activities that occur in the scene for an intelligent visual surveillance system. The model should handle motion patterns in a large scale. A crowd motion can be viewed as a dynamic system which contains a large number of local motions, and each local motion has its own dynamic system. When there are interactions among local motions exist, the global crowd motion, which consists of the local motions, is complicated.

The crowd model presented in this work is that we are trying to model the degree of similarity between the trained data and test data. This arrives from the fact that crowd motions are difficult to treat semantically. We learn crowd motion patterns using a CC-CTM model[10] by analyzing a large collection of crowd videos in an off-line manner. Then the structured and unstructured crowd motions in a new video can be represented semantically with the learned model. Our model is an analog of the topic model in text analysis, where texts are recognized based on the fact that same texts also share similar features. Documents in this work correspond to video clips, and topics correspond to crowd motion patterns. Crowd activities can be classified with the idea borrowing from text analysis. Our model can deal with multi-modality in crowd motions. Moreover, it is able to model correlations among different motion patterns. In addition, it models correlations among crowd motion patterns which is also

desirable. To classify activities, we use the large-margin training method to train classifiers[11].

2. Related Work

Human motion analysis is an active area in computer vision [12,13,14]. As one topic of human motion analysis, crowd scene analysis has attracted much research attention. An assumption of human motion analysis is that the bounding boxes of human bodies have already been detected (using either tracking or non-tracking methods). Knowning the bounding box information, features are extracted from local human body areas to represent the corresponding motions of people. The classification of human activities is then achieved by either finding good matches with priori known templates or learning a discriminant model.

Some methods for activity recognition of crowds rely on human trajectories obtained by various trackers [1,2,3,4,15,16,17]. For example, Alexei et al. [16] represent human motions as tacks of feature points, i.e. the trajectories. However, tracking of feature points can fail by occlusions or shadows, which restricts the usage such approaches to the analysis of complex activities in real world.

Low-level features and their statistics have been applied to describe human activities[18-19]. A well known technique is the scale-invariant spatio-temporal descriptors like[20], which describes irregular patterns of human motions using an ensemble approach without tracking interest points. The extracted features can directly be used to train discrimiant models for activity classification[7-8,21]. However, the lack of tracking information makes it difficult to distinguish different activities occuring simultaneously.

Topic models have been successfully used in generating semantic activity patterns from low-level feature co-occurrences. Wang et al.[24] used location and optical flow features along with hierarchical Bayesian approach to model activities and interactions. Li et al. [25] used spatiotemporal features along with a hierarchical pLSA to learn global behavior correlations. The method we propose here follows the idea in [8, 24]. The same datasets are used in this paper and [7], with totally different methods to classify the crowd activities. Semantic features are used in this paper and [24], with different topic models.

In this paper, we model crowd scenes with CC-CTM. CC-CTM enables us to process and extract semantic features from low level features including crowd spatial locations, motion direction and velocity without detecting and tracking specific motion objects. Multiple motion patterns may occur at different spatial locations in the scene with certain probabilities. Each of these motion patterns can be represented as high level information in our CC-CTM. The CC-CTM is elegant to model crowd activities. Our experiments also show that the CC-CTM offers an effective way to handle multi-modality and correlations within human motions. Different from topic models based on LDA (latent Dirichlet allocation)[22] or pLSA (probabilistic latent semantic analysis)[23], which assume that the topics are independent from each other, our model captures the correlations among topics by introducing a logistic normal prior (different from the Dirichlet prior in LDA) and a covariance matrix of topics.

The main difference between our model and the original latent CTM is that the number of topics is defined and the latent variables are guided by the classified datum during the sampling stage in the training pipeline. We show in this paper that class labels of training data are pushed directly into our model and, consequently, achieve much better performance in the analysis stage.

3. Topic Model for Crowd Activity

The fundamental principle of topic model is that each document is represented by a probabilistic distribution over topics, and each topic represents a probabilistic distribution over words. CC-CTM is a hierarchical model of video document collections. In this work, we propose CC-CTM for modeling crowd motions in surveillance scene. Our model provides an efficient way to capture crowd activities presented as multi-modality of motion patterns and the correlations among them.

3.1 Category Constrained Correlated Topic Model

The elements of CC-CTM and their conditional dependencies are depicted in the graphical model in **Fig. 2**. In this figure, shaded variables represent the observed variables, while unshaded variables represent the latent variables. Edges encode the conditional dependencies of the generative process. We now declare some terminologies used in this paper.

First, we represent an optical vector generated by vector quantization of spatio-temporal interest points as a single word. A video clip can then be represented by a "document" of many words. Second, our model is trained in a supervised manner using annotated labels, which significantly simplifies the training of model parameters and boosts recognition accuracy. The CC-CTM allows each document to contain multiple topics with different proportions. It can thus capture the multi-modal activities in crowd scene which typically contains multiple latent motion patterns.

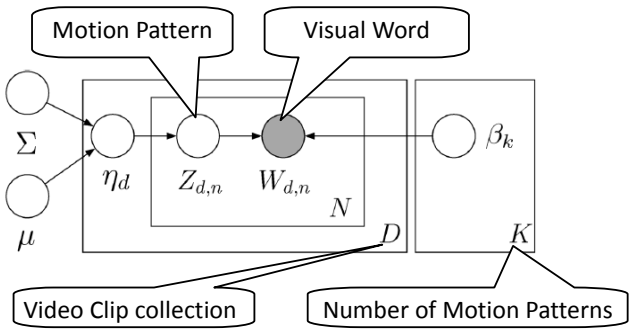


Fig. 2. Graphical Model of CC-CTM

Vocabulary. The vocabulary in this work is defined as the cartesian product of the location, motion direction, and motion magnitude word spaces, leading to a total of V words. This results in a high dimensional vocabulary.

Visual of Words. A *word* is the basic unit of vocabulary, which is defined as a word in a vocabulary indexed by $w_1, w_2, \cdots, w_V$. We represent words using unit-basis vectors which have a single component equal to one and all other components equal to zero. Using superscripts to denote components, the *v*-th word in the vocabulary is represented by a vector *w* such that $w_v = 1$ and other components equal to 0.

Documents. The documents are built by dividing the videos into short video clips, and then counting for each document *d* the number of times $w_{d,n}$ a word *w* occurs in it to obtain the bag-of-words representation of the document. The only observable random variables are

words $w_{d,n}$ mapping to visual words in this work, which correspond to low-level motion features, that is, quantized flow vectors and locations. Since a document *d* represents a video clip, the corpus D in CC-CTM can thus be mapped to the collection of video clips. Let $w_{d,n}$ denote the n-th visual word in the *d*-th video clip, which is an element in a *V* -term vocabulary of visual words.

Topics. A topic is mapped to a motion pattern. Let β denote a distribution over the vocabulary. Then the scene content can be viewed as a point on the $V-1$ simplex. The model contains *K* motion patterns (topics) $\beta_{1:k}$.

Topic assignments. Each visual word is assumed to be drawn from one of the *K* motion patterns. The motion pattern assignment $z_{d,n}$ is associated with the n-th visual word and the *d*-th video clip.

Topic proportions. Each video clip (document) is associated with a set of motion pattern proportions θ.As a result, θ_i is a distribution over motion pattern indices, and it reflects the probabilities with which visual words are drawn from each motion pattern in the collection. A natural parameterization of this multinomial is $\eta = \log(\theta_i / \theta_K)$.

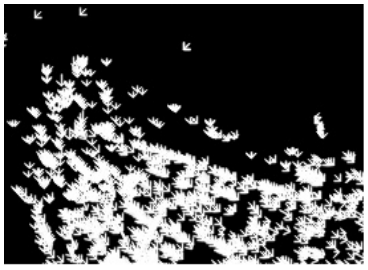
3.2 Visual words

To discover the semantic representation of crowd activity using CC-CTM, we need to define the vocabulary to build video documents. In our paper, a video document is a motion field represented with bag-of-visual-words model. The motion field is obtained by first using an existing optical flow method[26] to compute sparse optical flow vectors in each frame, and then combining the optical flow vectors from a temporal window of frames of the video into a single global motion field. In this paper, each clip contains 10 frames. A visual word is constructed by three types of information: crowd location, motion direction, and motion magnitude. The vector of feature *i* at time *t* is denoted by $V_i = (x_i, y_i, A_i, M_i)$. The examples can be seen in **Fig. 3**.

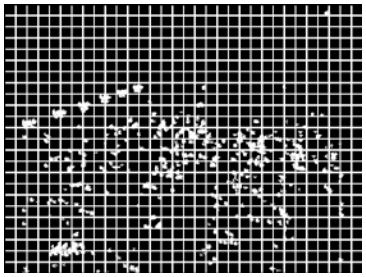
(a.1) scene 1 videio clip

(b.1) scene 2 videio clip

(a.2) scene 1 optical flow in the marked region

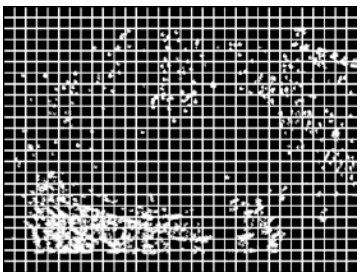


(b.2) scene 2 optical flow in the marked region



(a.3) location of 32×24 cell in visual words

(a) crowd scene1

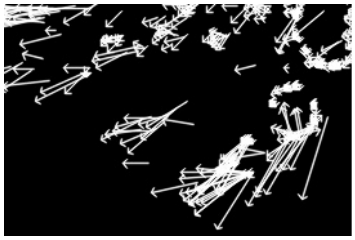


(b.3) location of 32×24 cell in visual words

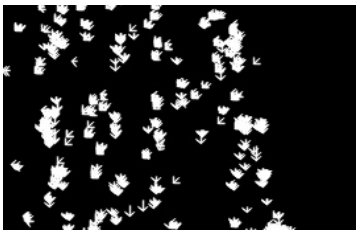
(b) crowd scene2

(c.1) scene 3 videio clip

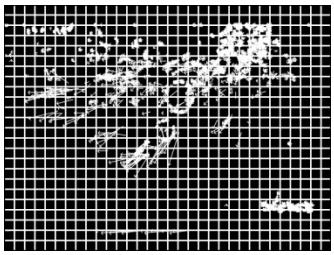
(d.1) scene 4 videio clip



(c.2) scene c optical flow in the marked region

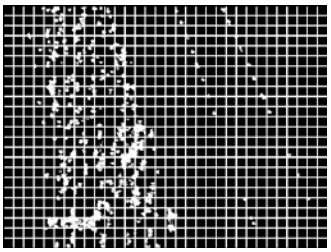


(d.2) scene 4 optical flow in the marked region



(c.3) location of 32×24 cell in visual words

(c) crowd scene3



(d.3) location of 32×24 cell in visual words

(d) crowd scene4

Fig. 3. Examples of crowd scenes (from UCF_CrowdsDataset [1]) and optical flows extracted from the highlighted regions.

Location (x_i, y_i). In surveillance videos, most of the activities are characteristic by the place where they occur. Thus, locations have to be taken into account when building the vocabulary. In crowded scenes, one pixel cannot be fully represented as crowd activities because of the existence of serious inter-object occlusions, small sized objects, similar appearance etc.. Thus, a pixel position (x_i, y_i) is quantized into non-overlapping cells of

20×20. Consequently,, for a video with a dimension of 640×480, we obtain a set of 32×24 cells as that shown in **Fig. 3**.

Motion direction (A_i). The moving pixels are filtered by thresholding the magnitude of the optical flow vectors. Moving pixels are further differentiated by quantizing their motion directions into labels from 0 to 7. (shown in **Fig. 4**) according to the angle intervals: $(-\frac{\pi}{8},\frac{\pi}{8}]$, $(\frac{\pi}{8},\frac{3\pi}{8}]$, $(\frac{3\pi}{8},\frac{5\pi}{8}]$, $(\frac{5\pi}{8},\frac{7\pi}{8}]$, $(\frac{7\pi}{8},\frac{9\pi}{8}]$, $(\frac{9\pi}{8},\frac{11\pi}{8}]$, $(\frac{11\pi}{8},\frac{13\pi}{8}]$, $(\frac{13\pi}{8},\frac{15\pi}{8}]$. Thus we have 8 possible motion directions in total.

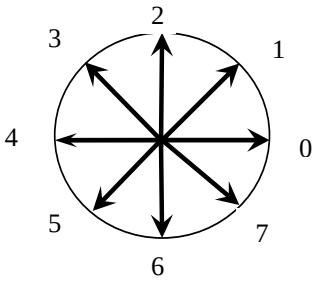


Fig. 4. Motion directions

Motion magnitude (M_i). To further characterize motion patterns, we associate motion magnitude with each detected interest point. In the video dataset, the motion object can be roughly classified into two categories based on its motion speed. For example, the motion magnitudes of detected interest points of objects with different motion categories are variant. Therefore, we apply a simple K-means clustering on the motion magnitudes with $K=2$.

3.3 Generative Process

The CC-CTM in this paper assumes that *N*-visual words from a video clip(document) *i* are created from the following generative process. Given topics $\beta_{1:K}$ (a distribution over the vocabulary), a *K*-vector μ and a $K\times K$ covariance matrix Σ, where μ and Σ are the mean and covariance of the normal distribution. The generative process is as below:

step 1: Randomly draw a *k*–dimensional vector $\eta_d \sim N(\mu,\Sigma)$ which determines the distribution of *K* types of crowd activity patterns in a video clip d.

step 2: For each visual word in the video clip d, $n\in\{1,\cdots,N_d\}$:

step 2.1: Choose a behavior $Z_{d,n}\,|\,\eta_d \sim \mathrm{Mult}\left(\frac{\exp(\eta)}{\sum_i \eta_i}\right)$ under the constraints of activity category k_d, where $Z_{d,n}$ is a K-dimensional unit vector with $Z_{d,n}=1$ indicating the k_d behavior is selected.

step 2.2: Choose a low-level motion feature $W_{d,n}\,|\,\{Z_{d,n},\beta_{1:K}\}$ from $\mathrm{Mult}(\beta_{Zd,n})$, where β is a distribution over the vocabulary of motion words under the indirect constraints of activity category k_d.

In correlated topic models, $\mu_{K\times 1}$, $\sum_{K\times K}$ and $\beta_{K\times V}$ are model parameters, while $\eta_{D\times K}$ and z are hidden variables. As a variational distribution $q(\cdot)$, we use a fully factorized model, where all variables are independently governed by a different distribution $q(\eta, z|\lambda, \nu, \phi) = q(\eta|\lambda, \nu,)q(z|\phi)$. Here $\lambda_{D\times K}$, $\nu_{D\times K}$ and ϕ are variational parameters. The only assumption for the variational inference is that η and z are independent and we do not specify any distributions for these hidden variables.

Given a collection of topics and the initial distribution over topic proportions $\{\hat{\beta}_{1:K}, \hat{\mu}, \hat{\Sigma}\}$, the parameters in the model are initialized as below:

$$\hat{\beta}_i \propto \sum_d \phi_{d,i} n_d \tag{1}$$

$$\hat{\mu} = \frac{1}{D}\sum_d \lambda_d \tag{2}$$

$$\hat{\Sigma} = \frac{1}{D}\sum_d I\upsilon_d^2 + (\lambda_d - \hat{\mu})(\lambda_d - \hat{\mu})^T \tag{3}$$

where n_d is the vector of word counts for document d.

The log probability of one video clip $w = \{w_1, w_2, \cdots, w_N\}$, which is defined as follows:

$$P(d|\mu, \Sigma, \beta, k_d) = \int P(\eta|\mu, \Sigma)\left(\prod_{n=1}^{N} P(z_n|\eta, k_d) P(w_n \mid z_n, \beta)\right) d\eta \tag{4}$$

The log probability of the video clips collection (document corpus D) is

$$P(D|\mu, \Sigma, \beta, k_d) = \prod_{d=1}^{M}\left(\int P(\eta|\mu, \Sigma)\left(\prod_{n=1}^{N} P(z_n|\eta, k_d) P(w_n|z_n, \beta)\right) d\eta\right) \tag{5}$$

In order to perform parameters estimation for CC-CTM model, We use a collection of training video documents and adopt the variational expectation maximization (EM) algorithm proposed in CTM[10]. We refer reader to this reference for further details on the parameter estimation algorithm.

4. Modeling Multi-modal of Crowd Activities

The generative model we proposed in this work can model the dynamics and complex scenes by capturing spatial, directional and velocity's dependencies among different motion patterns in an unsupervised framework. The multi-modality of crowd activities is represented in high level motion features, which is a statistical model built from low level features. The process of our method is illustrated in **Fig. 5**.

Crowd activities recognition in this paper is processed in two stages, ie. the learning stage and the recognition stage. Low-level motion features are extracted from optical flow field for the training video datasets. The feature of each optical flow vector is mapped to a visual word which is used to construct the visual dictionary. A document in corpus D is represented with bag-of-visual-words model. In this paper, CC-CTM is used to model the dynamics of crowd by capturing spatial-temporal dependencies between different behaviors under a semi-supervised learning way for the same scene. A crowd activity presenting as a crowd motion pattern, is represented as the semantic feature referenced from CC-CTM.

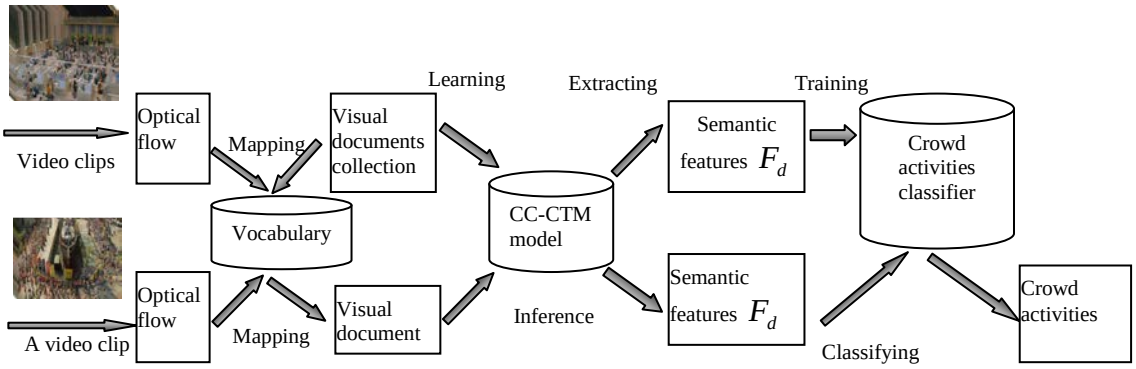


Fig. 5. Process of crowd activities recognition.

4.1 Crowd Activity Modeling with CC-CTM

CC-CTM model can be used to infer the crowd activity in a video clip after learning from training examples. The biggest difference between CC-CTM and the original topic models appears in learning stage. In the original topic models, such as LDA or CTM, we can only extract the visual words $w_i(i=1,2,\cdots N)$ in each video clip, but the topic keeps latent and we do not know which topic z_k is assigned for the word w_i. In this paper, our aim is to recognize crowd activities. All training videos can be classified and labeled according to the activities happening in them. Hence this important information can be used in training stage under the semi-supervised learning framework.

CC-CTM is a modified version of the CTM model. At the beginning of learning stage, all parameters of the CC-CTM model are initialized as original CTM model, except for the topic assignments $z_{d,n}$ (the n-th visual word and of the *d*-th video clip), i.e. we have *K* seed documents if the training datasets contain K categories. All visual words $w_i(i=1,2,\cdots N)$ in a seed document are enforced to assign a topic k according to the document label associated with the activity category.

Parameters of the CC-CTM model are estimated during training. The topic proportions of an input video clip can be inferred from the trained model, which can be represented as semantic feature in a more compact representation (shown in **Fig. 6**).

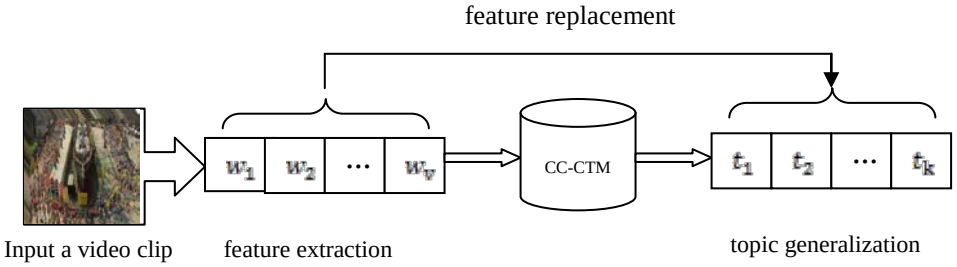


Fig. 6. feature replacement with semantic feature.

4.2 Crowd behavior classification

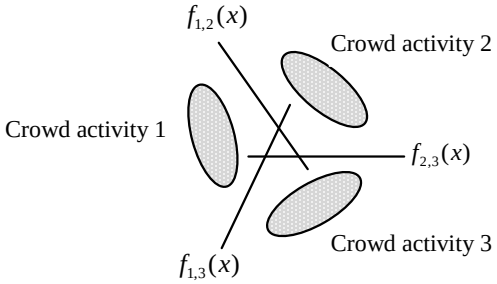


Fig. 7. One-against-one classifier.

The high level feature vector for the a video clip (document) is defined as

$$F_d = [t_1, t_2, \cdots, t_K] \tag{6}$$

where $t_i (i = 1,2, \cdots\cdots, K)$ is the distribution of the *i*-th class motion pattern in this model. In this study, each crowd activity is represented by a mixture distribution of K types of motion pattern, which is defined as feature vector F_d in Eq.(6). To predict crowd activities, a multi-class SVM classifier is trained in an one-against-one manner. Specifically, we use the off-the-shelf LIBSVM toolbox[27] to train our model with the RBF kernel, which is defined as:

$$K(x_i, x_j) = \exp(-\gamma \| x_i - x_j \|^2), \qquad for \ \ \gamma > 0 \tag{7}$$

Where x_i and x_j are the feature vectors, and γ is the penalty factor .

The cross-validation approach is employed to train the proposed classifier. In cross-validation, all training samples are partitioned into N subsets of equal size. N-1 subsets are used for training and the left one subset is used for testing. To solve the noise-borne problem, we optimize the slack variable and penalty factor γ.

5. Experiments

In this work, the video datasets are clipped into 10 frames evenly without intersection. We test our algorithm on 2 datasets: UCF_CrowdsDataset and Crowd_PETS09. Four videos of crowded scenes selected from UCF_CrowdsDataset are clipped into video clips in 10 frame evenly with the resolution of 640×480 pixels. The video clip samples and their part optical flow are shown in **Fig. 3**. The crowded motion pattern in these selected vides are quite differently from each other. People in scene 1 are in unstructured motion pattern with random smallscale movement. People in scene 2 are in structured motion pattern with small-scale movement in a circular way. People in scene 3 are in unstructured motion pattern with large-scale movement. People in scene 4 are in structured motion pattern with small-scale movement in line. CC-CTM model provides a D by K matrix of the variational mean parameter for each document's topic proportions.

We select 20 training video clips for each scene. Each video clip is represented as a document succinctly composed by visual code words after extracting low-level optical flow feature. The number of visual code words in vocabulary is $32\times24\times8\times2=12288$. The low-level feature in the video clip is in high dimensional. In order to recognize these 4 distinct scenes, the training datasets are classified into 4 categories, i.e the topic number K is 4. All of the visual words $w_i(\mathrm{i}=1,2,\cdots \mathrm{N})$ in the seed document are enforced to assigned topic k according to the document label associated with activity category. The topic (activity) proportion feature of each video clip is inferred from the learned CC-CTM model. The cluster results (shown in **Fig. 8**) in 4 clusters apparently demonstrate that the high-level feature in low-dimension efficiently represents the crowd activity, and the same scene features are tightly clustered.

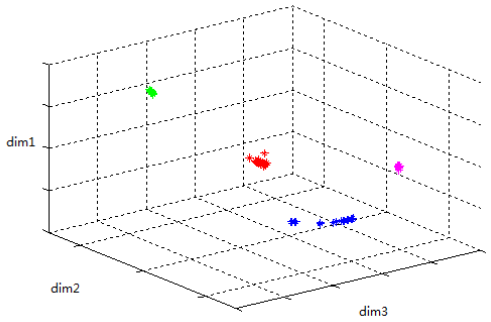


Fig. 8. Cluster results for semantic features.

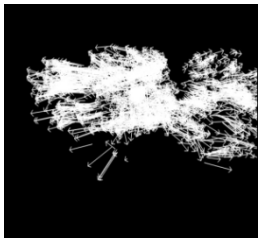
We do further experiment on recognition different crowded activities in the same surveillance scene. There are 8 classes of activities in Crowd_PETS09 (shown in **Fig. 9**).

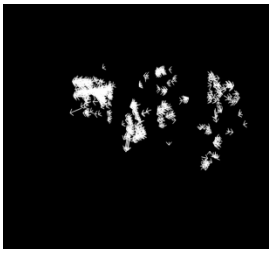

(a. 1) Image sequences

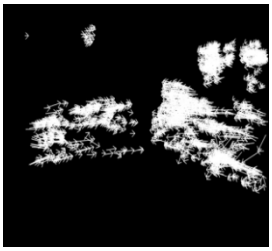

(b. 1) Image sequences

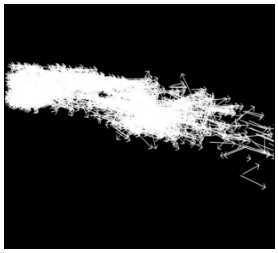

(c. 1) Image sequences


(d. 1) Image sequences


(a. 2) Optical flow field


(b. 2) Optical flow field


(c. 2) Optical flow field


(d. 2) Optical flow field

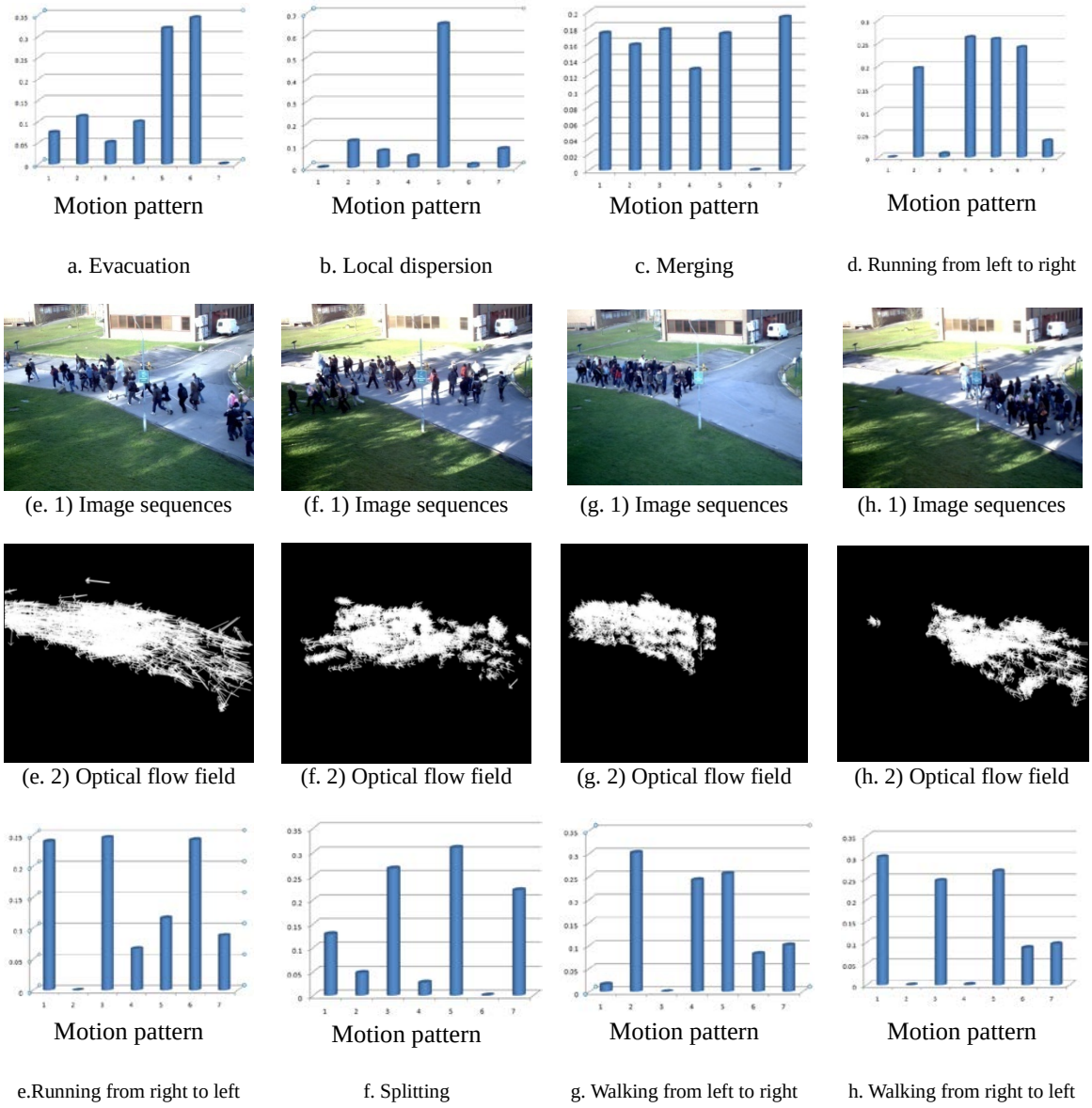


Fig. 9. 8 classes of crowd activities and motion patterns.

In this paper, we process event recognition sequences which are organized in the datasets from S0 to S3 under the same camera position of "View_001". There are 2000 documents in total and every adjacent 10 frames is organized as a video clip with a 640×480 resolution. The dataset are split into a training set and a testing set. We obtained good experimental results when the samples for each activity are evenly selected in the training data. The confusion matrices for activity classification with our topic model are shown by **Fig. 10** and **Fig. 11** respectively.

	Evacuation	Local dispersion	Merging	Running from left to right	Running from right to left	Splitting	Walking from left to right	Walking from right to left
Evacuation	100%						20.59%	
Local dispersion		100%	4.0%				2.94%	
Merging			72%				14.71%	15.26%
Running from left to right				100%			2.94%	
Running from right to left					100%			
Splitting			4.0%			100%		11.54%
Walking from left to right			8.0%				58.82%	
Walking from right to left			12.0%					73.2%

Fig. 10. Confusion matrices obtained with our CC-CTM model.

	Evacuation	Local dispersion	Merging	Running from left to right	Running from right to left	Splitting	Walking from left to right	Walking from right to left
Evacuation	100%		8%					11.54%
Local dispersion		71.43%					8.82%	11.54%
Merging			76%				8.82%	15.38%
Running from left to right				100%			2.94%	
Running from right to left					100%	14.28%		
Splitting		28.57				85.72%		3.85%
Walking from left to right			16%				79.41%	
Walking from right to left								57.7%

Fig. 11. Confusion matrices obtained with original CTM model.

The recognition results of "Evacuation", "Local dispersion", "Running" and "Splitting" are perfect. The main reason is that the optical flow features for these activity categories are highly discriminative. Apparently the accuracy of the "Walking" activity is much lower than other classes.

Method in paper [7] used a direction model to represent the motion pattern and to detect the event on the Crowd_PETS09 s3 dataset including 1000 frames. Benabbas [7] proposed two classifiers, one for detecting motion-speed-related events and the second for detecting crowd convergence and divergence events. The method depends on the low-level features, and the recognition accuracy is high for the crowd activity with apparently different motion feature. The comparison results are show in **Table 1**. Results by our method are obviously better than the results in [7] except for the walking activity. Our method also performs better than the original CTM, mainly due to the usage of the semi-supervised learning approach.

Table 1. Classification accuracy by three different methods.

Methods / Event types	Recognition accuracy with our method	Recognition accuracy with original CTM	Recognition accuracy with method in paper[7]
Evacuation	100%	100%	100%
Local dispersion	100%	71%	49%
Merging	72%	76%	46%
Running	100%	100%	92%
Splitting	100%	86%	99%
Walking	66%	69%	68%
Average Accuracy	90%	84%	76%

6. Conclusion

In this paper we have presented a novel approach to analyze crowd activities using CC-CTM. It bypasses time-consuming methods such as background subtraction and person detection and rather resorts to global motion information obtained from optical flow vectors to model the motion magnitude and velocity at each spatial location of the scene. With predefined number of motion patterns, these models use the distributions of mixture activity patterns to analyze the main crowd activities. We evaluate the performance of our approach on multiple datasets. These experiments show that our approach is applicable to a wide range of scenes which consist of low and high crowd density scenes as well as structured and unstructured scenes. In addition, the method presented in this paper detects groups of people even in the presence of occlusions, which facilitates the detection of group-related events such as merging and splitting.

There are four main contributions proposed in this paper. First, we propose a novel bag-of-words representation for low level motion features in video, which avoid the difficulties of severe inter-object occlusion, small object size, similar appearance, etc. for object tracking in intelligent surveillance systems. Second, since the video datasets are very different from text datasets, we introduce a Category Constrained Correlated Topic Model (CC-CTM), an improved topic model, in the study area of video computation and understanding. The topic models are intend to explore text mining traditionally. Third, CC-CTM can naturally explore crowd activities in high level features with activities correlated semantic representation in a video clip. Fourth, experimental results under different condition presented in this paper to show that the proposed models achieve 8 kinds of crowd activities recognition accuracies on PETS benchmark datasets .

In future, we plan to address some specific problems in order to improve the results. On one hand, it is urgent to develop a better optical flow algorithm for highly crowded scene. On the other hand, we plan to improve the CC-CTM model to detect the abnormal event, which is more desirable for the public security in the intelligent surveillance system.

References

[1] Ali S, Shah M., "Floor fields for tracking in high density crowd scenes[C]," *Computer Vision–ECCV 2008*, Springer Berlin Heidelberg, 1-14, 2008. Article (CrossRef Link).

[2] Rodriguez M, Ali S, Kanade T., "Tracking in unstructured crowded scenes[C]," *Computer Vision*, in *Proc. of 2009 IEEE 12th International Conference on. IEEE*, 1389-1396, 2009. Article (CrossRef Link).

[3] Bera A, Manocha D., "Realtime multilevel crowd tracking using reciprocal velocity obstacles[C]," *Pattern Recognition (ICPR)*, in *Proc. of 2014 22nd International Conference on. IEEE*, 4164-4169, 2014. Article (CrossRef Link).

[4] Fradi H, Dugelay J L., "Spatial and temporal variations of feature tracks for crowd behavior analysis[J]," *Journal on Multimodal User Interfaces*, 1-11, 2015. Article (CrossRef Link).

[5] Shao J, Kang K, Loy C C, et al., "Deeply learned attributes for crowded scene understanding[C]," in *Proc. of CVPR*, 4657-4666, 2015. Article (CrossRef Link).

[6] Wu S, Moore B E, Shah M., "Chaotic invariants of lagrangian particle trajectories for anomaly detection in crowded scenes[C]," *Computer Vision and Pattern Recognition (CVPR)*, in *Proc. of 2010 IEEE Conference on. IEEE*, 2054-2060, 2010. Article (CrossRef Link).

[7] Benabbas Y, Ihaddadene N, Djeraba C., "Motion pattern extraction and event detection for automatic visual surveillance[J]," *Journal on Image and Video Processing*, 7, 2011. Article (CrossRef Link).

[8] Rodriguez M, Sivic J, Laptev I, et al., "Data-driven crowd analysis in videos[C]," *Computer Vision (ICCV)*, in *Proc. of 2011 IEEE International Conference on. IEEE*, 1235-1242, 2011. Article (CrossRef Link).

[9] YI S, WANG X, LU C et al., "L0 regularized stationary time estimation for crowd group analysis[C]," *CVPR '14*, 2219-2226, 2014. Article (CrossRef Link).

[10] Blei D M, Lafferty J D., "A correlated topic model of science[J]," *The Annals of Applied Statistics*,1(1): 17-35, 2007. Article (CrossRef Link).

[11] Meyer D., "Support Vector Machines The Interface to libsvm in package e1071[J]," *R News*, 1(5):1—3, 2001. Article (CrossRef Link).

[12] Popoola O P, Wang K., "Video-Based Abnormal Human Behavior Recognition—A Review[J]," *IEEE TRANSACTIONS ON SYSTEMS, MAN, AND CYBERNETICS—PART C: APPLICATIONS AND REVIEWS, NOVEMBER 2012*, 42(6)：865-878, 2012. Article (CrossRef Link).

[13] Li T, Chang H, Wang M et al., "Crowded scene analysis: A survey[J]," *Circuits and Systems for Video Technology, IEEE Transactions on*, 25(3): 367-386, 2015. Article (CrossRef Link).

[14] Raghavendra R, Cristani M, Bue A D et al., "Anomaly Detection in Crowded Scenes: A Novel Framework Based on Swarm Optimization and Social Force Modeling[M]," *Modeling, Simulation and Visual Analysis of Crowds*, Springer New York, 2013. Article (CrossRef Link).

[15] Fradi H, Dugelay J L., "Sparse Feature Tracking for Crowd Change Detection and Event Recognition[C]," *Pattern Recognition (ICPR)*, in *Proc. of 2014 22nd International Conference on. IEEE*, 4116-4121, 2014. Article (CrossRef Link).

[16] Alexei Gritai, Yaser Sheikh & Mubarak Shah, "On the use of Anthropometry in the Invariant Analysis of Human Actions [A]," in *Proc. of IEEE Proceedings of Intelligent Conference on Pattern Recognition*, 923-926, 2004. Article (CrossRef Link).

[17] Wang X, Ma K T, Ng G W et al., "Trajectory analysis and semantic region modeling using nonparametric hierarchical bayesian models[J]," *International journal of computer vision*, 95(3): 287-312, 2011. Article (CrossRef Link).

[18] Varadarajan J, Odobez J M., "Topic models for scene analysis and abnormality detection[C]," *Computer Vision Workshops (ICCV Workshops)*, in *Proc. of 2009 IEEE 12th International Conference on*, *IEEE*, 1338-1345, 2009. Article (CrossRef Link).
[19] T. Xiang and S. Gong, "Video behavior profiling for anomaly detection[J]," *IEEE Trans. Pattern Anal. Mach. Intell*, 30(5):893-908, May 2008. Article (CrossRef Link).
[20] O. Boiman andM. Irani, "Detecting irregularities in images and in video[C]," in *Proc. of 10th IEEE Int. Conf. Comput. Vision*, vol. 1:462–469, 2005. Article (CrossRef Link).
[21] Hasan S, Ukkusuri S V, "Urban activity pattern classification using topic models from online geo-location data[J]," *Transportation Research Part C: Emerging Technologies*, 44: 363-381, 2014. Article (CrossRef Link).
[22] Blei D M, Ng A Y, Jordan M I, "Latent dirichlet allocation[J]," *the Journal of machine Learning research*, 3: 993-1022, 2003. Article (CrossRef Link).
[23] Hofmann T., "Probabilistic latent semantic indexing[C]," in *Proc. of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval. ACM*, 50-57, 1999. Article (CrossRef Link).
[24] X. Wang, X. Ma, and W. E. L., "Grimson. Unsupervised activity perception in crowded and complicated scenes using hierarchical bayesian models[J]," *IEEE Trans. Pattern Anal.Mach. Intell*, 31(3):539-555, Mar 2009. Article (CrossRef Link).
[25] J. Li, S. Gong and T. Xiang, "Global behavior inference using probabilistic latent semantic analysis[C]," in *Proc. of 2009 IEEE 12th International Conference on Computer Vision Workshops*, *ICCV Workshops*, 1338-1345, 2008. Article (CrossRef Link).
[26] Brox T., "Optical Flow[M]," *Computer Vision*, Springer US, 565-569, 2014. Article (CrossRef Link).
[27] CHANG C-C,LIN C-J., "Libsvm: A library for support vector machines[J]," *ACM Transactions on Intelligent Systems and Technology (TIST)*, 2(3):27, 2011. Article (CrossRef Link).
[28] PETS 2009 Benchmark Data. http://www.cvg.reading.ac.uk/PETS2009/a.html. Article (CrossRef Link).

Xianping Huang received the Ph. D. degree in Control Theory and Control Engineering from Zhejiang University of Technology, Hangzhou, China in 2015. Currently, she is an Associate Professor in college of computer science & technology, Zhejiang University of Technology, China. Her research interests include computer vision, computer graphics and visualization.

Wanliang Wang received Ph.D. degree in Control Theory and Control Engineering in 2001. He has devoted nearly 20 years to educational work and now he is the dean of College of Computer Science and Technology in Zhejiang University of Technology. As a researcher, he is leading a large research group in the field of simulation for small hydropower projects and many valuable achievements have made. His research interests cover intelligent algorithms and network control.

Guojiang Shen received the Ph.D. degree in Control Science and Engineering from Zhejiang University, Hangzhou, China, in 2004. He is currently a Professor in College of Computer Science and Technology, Zhejiang University of Technology. His current research interests include intelligent control theory and application, advanced control technology andapplication, and urban road traffic modeling and controltechnology.

Xiaoqing Feng receivedthe Ph.D.degree in computer science in computer science and technology from Zhejiang University, Hangzhou, China in 2010.Currently, she is an Associate Professor in School of information,Zhejiang University of Finance & Economics, China. Her research interests include multimedia signal processing and information security.

Xiangjie Kong received the Ph. D. degree in Control Science and Engineering from Zhejiang University, Hangzhou, China, in 2009. Currently, he is an Associate Professor in School of software, Dalian University of technology, China. He has served as Editor Board Member of SpringerPlus, Guest Editor of several international journals,Workshop Chair or PC Member of a number of conferences. Dr. Kong has published over 30 scientific papers in international journals and conferences (with 20+ indexed by ISI SCIE). His research interests include big traffic data, social computing, and mobile computing. He is a Member of IEEE and ACM, a Senior Member of CCF.