Gao GY, Ma HD. Movie scene recognition using panoramic frame and representative feature patches. JOURNAL OF COMPUTER SCIENCE AND TECHNOLOGY 29(1): 155–164 Jan. 2014. DOI 10.1007/s11390-013-1418-1

Movie Scene Recognition Using Panoramic Frame and Representative Feature Patches

Guang-Yu Gao[1,2] (高广宇) and Hua-Dong Ma[1,*] (马华东)

[1] *Beijing Key Laboratory of Intelligent Telecommunications Software and Multimedia, Beijing University of Posts and Telecommunications, Beijing 100876, China*

[2] *School of Software, Beijing Institute of Technology, Beijing 100081, China*

E-mail: guangyu.ryan@gmail.com; mhd@bupt.edu.cn

Received December 27, 2012; revised November 26, 2013.

Abstract Recognizing scene information in images or videos, such as locating the objects and answering "Where am I?", has attracted much attention in computer vision research field. Many existing scene recognition methods focus on static images, and cannot achieve satisfactory results on videos which contain more complex scenes features than images. In this paper, we propose a robust movie scene recognition approach based on panoramic frame and representative feature patch. More specifically, the movie is first efficiently segmented into video shots and scenes. Secondly, we introduce a novel key-frame extraction method using panoramic frame and also a local feature extraction process is applied to get the representative feature patches (RFPs) in each video shot. Thirdly, a Latent Dirichlet Allocation (LDA) based recognition model is trained to recognize the scene within each individual video scene clip. The correlations between video clips are considered to enhance the recognition performance. When our proposed approach is implemented to recognize the scene in realistic movies, the experimental results shows that it can achieve satisfactory performance.

Keywords movie scene recognition, key-frame extraction, representative feature, panoramic frame

1 Introduction

Over the last decade, the number of digital images and videos has grown tremendously. At the same time, some important but difficult problems have been raised. It is one of the most challenging problems to know "Where am I" in the computer vision area, namely to recognize the place/scene. Scene recognition devotes to getting location categories, while place recognition always focuses on finding the exact location in realistic.

According to the type of input data (image or video), scene recognition can be classified into scene recognition for image and scene recognition for video. In the past decades, many studies focused on scene recognition for image[1–5]. Some of these studies (e.g., [3]) have achieved satisfactory performance in certain image datasets (e.g., [6]). For scene recognition for video, Marszałek *et al.*[7] proposed a context-based method to recognize the scene of videos, and Engels *et al.*[8] proposed an automatic annotation scheme for unique locations from videos. However, to the best of our knowledge, there are still few researches on the scene recognition for video. This is because videos contain more complex scenes than images. The complexity is caused by various changes, such as moving direction and lighting conditions. Moreover, well-annotated video scene datasets are lacked.

In this paper, we focus on challenging scene recognition for video, especially movies or teleplays, and propose a panoramic frame and representative feature patches based movie scene recognition approach (the overview is shown in Fig.1). More specifically, we first describe how to segment the whole movie into shots and video scenes. Then, we introduce a novel key-frame extraction method using panoramic frame and a representative feature extraction method. These methods can ensure that we obtain enough representative and robust features for scene recognition. Finally, a recognition model is designed to recognize the scene category of each video clip. Our main contributions are as follows.

Regular Paper

The research is supported by the National Funds for Distinguished Young Scientists of China under Grant No. 60925010, the Specialized Research Fund for the Doctoral Program of Higher Education of China under Grant No. 20120005130002, the Cosponsored Project of Beijing Committee of Education, the Funds for Creative Research Groups of China under Grant No. 61121001, and the Program for Changjiang Scholars and Innovative Research Team in University of China under Grant No. IRT1049.

[*] Corresponding Author

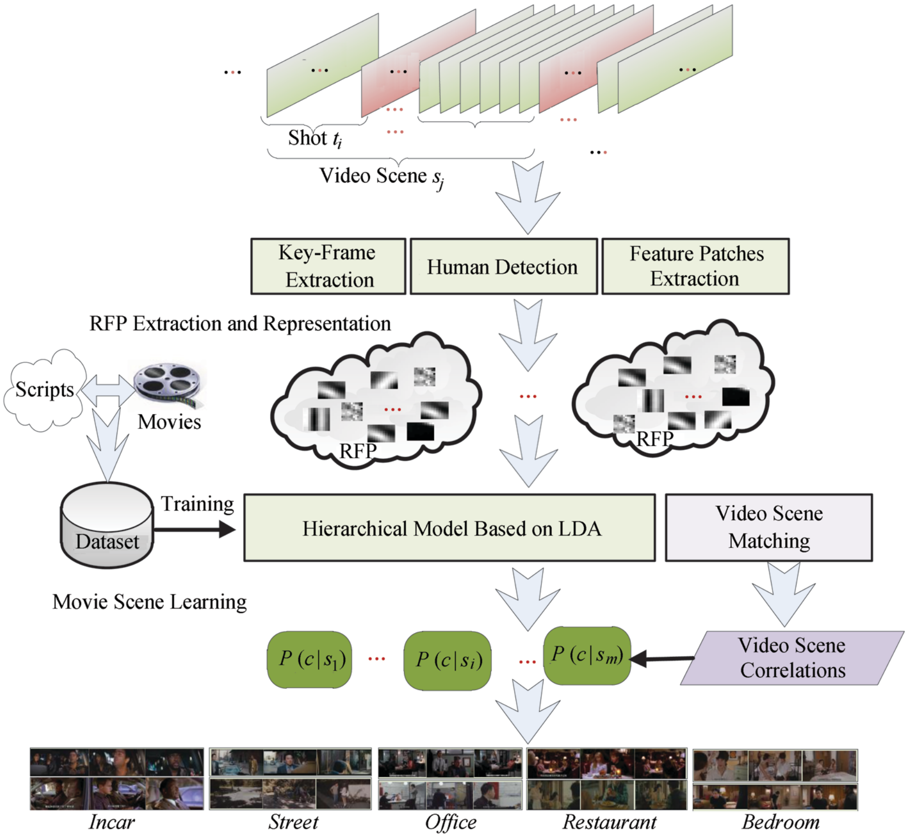


Fig.1. Overview of the proposed movie scene recognition approach. $P(c|s_i)$ is the output probability, which means the probability that s_i is classified to scene category c.

- We propose to extract panoramic frames as key-frames in each video shot through video registration.
- We introduce the representative feature patches (RFPs) as middle-level features to comprehensively represent each video scene. The RFPs are extracted in the previously obtained panoramic frames, where the regions of human are eliminated effectively.
- In the scene recognition stage, the informative correlations between video scenes are used to enhance the recognition performance of individual video scene.

The remainder of this paper is organized as follows. We review some related work in Section 2, and the overview of our approach is introduced in Section 3. We also present details of video segmentation in Section 4. Panoramic frame based key-frames and RFPs extraction are discussed in Section 5. In Section 6, we describe how to get an enhanced generative model for movie scene recognition using video scene correlations. Experimental results are shown in Section 7. Finally, we conclude this paper in Section 8.

2 Related Work

A scene is defined as a site where the image or video is taken. Examples of scenes include office, bedroom and so on. So far, there are many related scene recognition methods, and some of them perform well, especially for images. For example, a probabilistic neural network (PNN) was used for indoor versus outdoor scene classification[1]. Zhou *et al.*[9] presented a novel Gaussianized vector to represent the scene images for unsupervised recognition. Besides, Oliva and Torralba[10] incorporated the idea of using global frequency with local spatial constraints to recognize scenes.

However, scenes in the same category are presented in various forms of appearances, thus many researches thought that the local features would be more suitable to construct the scene model. For example, Liu *et al.*[4] utilized Maximization of Mutual Information (MMI) co-clustering approach to discover clusters of semantic concepts for scene modeling. Li *et al.*[2] proposed a hierarchical Bayesian to classify 13 image scenes. Moreover, Lazebnik *et al.*[6] argued that methods based on basic bag-of-features representation will do scene recognition better. Thus, they remained sympathetic to the goal of developing robust and geometrically invariant structural object representations, and proposed the spatial pyramid matching for recognizing natural scene categories. In addition, Wu and Rehg[3] introduced a new visual descriptor, named, Census Transform Histogram (CENTRIST), to recognize both topological places and scenes categories. Xiao *et al.*[11] even proposed the extensive Scene Understanding (SUN) database that contains 899 categories and 130 519 images for scene recognition.

Nevertheless, the aforementioned methods mainly handle the recognition for specially collected images, and cannot deal with the same problems well in movies. Recently, there are also several studies on movie scene recognition, such as [12-13]. Huang *et al.*[13] developed the scene categorization scheme using a Hidden Markov Model (HMM)-based classifier. However, these methods concentrate on simple videos only, i.e., basketball video, football video and commercial video, and cannot be applied to movies. Movie scene recognition is more challengeable since the physical location appearance is various in camera viewpoints, partial occlusion and lighting changes, etc. Schaffalitzky and Zisserman[14] described the progress in matching shots of the same 3D location in a film. But both their local invariant descriptors extraction and the shots' matching process are very time consuming. Meanwhile, an improved unsupervised classification method was proposed by Héritier *et al.*[15-16] to extract and link places features and cluster recurrent physical locations (key-places) within a movie. Their work focuses on near-duplicate detection which is composed of footage or images of the same object or same background but taken at different time and/or different places. Bosch *et al.*[17] also presented a pLSA-based scene classification method mainly for images, but they tested key-frames from a movie. In their movie, there are only a few images that could be accurately classified. That is to say it is still a difficult task to directly use image scene recognition method to do movie scene recognition.

Actually, Ni *et al.*[18] presented an efficient algorithm for recognizing the locations of a camera/robot in the learned environment using only the images it captures. However, the locations or scenes in movies are not so specific and stable. In fact, the most related work to our approach is the studies of [7-8]. Engels *et al.*[8] proposed the automatic annotation scheme for unique locations from videos. Although it is satisfactory and accurately to annotate video locations with location word, it is based on the hypothesis that the transcripts are available. Marszałek *et al.*[7] proposed a joint framework for action and scene recognition and also demonstrated the enhanced recognition for both of them in natural video. However, their main purpose is to do action recognition and their scene recognition results are not satisfactory.

In order to deal with the various appearances in movie scene recognition, we should extract more robust and representative feature patches and more efficient recognition model as well as use domain knowledge. Thus, in this paper, we use the panoramic frame as the key-frame, and choose the representative feature patches (RFPs) to represent each video clip. Meanwhile, local patches drops in human regions are excluded because human regions always mean the noisy for movie scene. In addition, we consider the reoccurred movie scenes referring to the same place as the optimal candidates for enhanced recognition.

3 Overview

As shown in Fig.1, the proposed method consists of five stages: 1) video segmentation; 2) key-frame extraction; 3) representative local features extraction; 4) LDA-based classifier construction; 5) video scene correlations based enhancement. At first, the video is segmented into shots and scenes, and this is introduced in Section 4. Then, the panoramic frame is obtained as the key-frame for each shot, and representative feature patches are extracted from these panoramic frames belonging to the same video scene, detailed in Section 5. After that, the Bayesian classifier with LDA model is trained to recognize the scene category for each video scene, and an enhanced scene recognition processing is implemented with near duplicated video scene detection. The final processes are described in Section 6.

4 Video Segmentation

In this section, video segmentation, including shot boundary detection and scene detection, is adopted to segment each movie into clips. While there are two different meanings for the word *scene* (one is used in scene recognition, the other is the description which refers to a group of shots in video composition), we maintain the meaning for *scene* in scene recognition, and use video scene (VSC) to represent the meaning in video composition definition for notational distinction. At first, an accelerating shot boundary detection method[19] is adopted to segment a movie into shots efficiently. After that, we propose a multi-modality movie scene detection method using kernel canonical correlation analysis based feature fusion[20]. Specifically, feature movies are often filmed in open and dynamic environments using moving cameras, and also have continuously changing contents. Thus, we focus on the association extraction of visual features $\boldsymbol{x}$ (e.g., color, gradient, motion) and audio features $\boldsymbol{y}$ (e.g., mel-frequency cepstral coefficients (MFCCs), short time energy log measure (STE)). Based on the Kernel Canonical Correlation Analysis (KCCA), all these features are fused for efficient VSC detection as follows.

Canonical correlation is to choose $\boldsymbol{w_x}$ and $\boldsymbol{w_y}$ to maximize the correlation between the two variables $\boldsymbol{x}$ and $\boldsymbol{y}$. Namely, the function to be maximized is

$$\rho = \max_{\boldsymbol{w_x},\boldsymbol{w_y}} \frac{\boldsymbol{w_x}^{\mathrm{T}}\boldsymbol{C_{xy}}\boldsymbol{w_y}}{\sqrt{\boldsymbol{w_x}^{\mathrm{T}}\boldsymbol{C_{xx}}\boldsymbol{w_x}\boldsymbol{w_y}^{\mathrm{T}}\boldsymbol{C_{yy}}\boldsymbol{w_y}}}, \tag{1}$$

where,

$$\boldsymbol{C}(\boldsymbol{x},\boldsymbol{y}) = E\left[\begin{pmatrix}\boldsymbol{x}\\ \boldsymbol{y}\end{pmatrix}\begin{pmatrix}\boldsymbol{x}\\ \boldsymbol{y}\end{pmatrix}^{\mathrm{T}}\right] = \begin{pmatrix}\boldsymbol{C_{xx}} & \boldsymbol{C_{xy}}\\ \boldsymbol{C_{yx}} & \boldsymbol{C_{yy}}\end{pmatrix}. \tag{2}$$

$\boldsymbol{C_{xx}} = E[\boldsymbol{x}\boldsymbol{x}^{\mathrm{T}}]$ and $\boldsymbol{C_{yy}} = E[\boldsymbol{y}\boldsymbol{y}^{\mathrm{T}}]$ are the auto-covariance matrix. $\boldsymbol{C_{xy}} = E[\boldsymbol{x}\boldsymbol{y}^{\mathrm{T}}]$ and $\boldsymbol{C_{yx}} = E[\boldsymbol{y}\boldsymbol{x}^{\mathrm{T}}]$ are the polled-covariance matrix. Thus, while $\boldsymbol{x}$ and $\boldsymbol{y}$ refer to the visual and audio features respectively, the resulting combined audio-visual feature vector is thus given by

$$\boldsymbol{\mathcal{Z}} = \begin{pmatrix}\boldsymbol{w_x}\\ \boldsymbol{w_y}\end{pmatrix}^{\mathrm{T}}\begin{pmatrix}\boldsymbol{x}\\ \boldsymbol{y}\end{pmatrix}. \tag{3}$$

After that, a similarity graph is constructed with spatial-temporal coherent shots and partitioned to generate the VSC boundaries.

5 Representative Features Patches Extraction

A VSC usually refers to a group of shots taken in the same physical location, thus single frame does not take sufficient information of it. Therefore, we get several key-frames in a VSC, and extract the representative features in these key-frames. Concretely, while the frames in a shot are captured by one time camera motion, we use the panoramic frame obtained by video (frame-to-frame) registration, as the key-frame. Then, in order to get more representative features, RFPs are extracted from key-frames.

5.1 Key-Frame Extraction

In order to get the most representative features as well as to reduce the noise involved in redundant frames, we propose a novel key-frame extraction method using panorama frame. Recently, several key-frame selection methods were proposed to choose an appropriate number of key-frames, especially for a dynamic shot with larger motion of actor or camera[21-22]. However, these methods may satisfy the requirement of applications such as video summary rather than movie scene recognition. That is because to compare with these applications, we need more comprehensive and representative features of a VSC to distinguish different scene categories. While the panoramic frame obtained by video frame registration contains more completed features, Xiao *et al.*[23] introduced the problem of scene viewpoint recognition with panoramic images organized into 26 place categories. Thus, in this paper we also use the panoramic frame to construct our key-frames.

The panoramic frame can be obtained by video registering. In this paper, in order to obtain a more robust panoramic frame, we adopt the RVR method proposed by Ghanem *et al.*[24] As described in [24], we get the panoramic frame by calculating the homography matrix between two frames, since with the homography matrix, we can map all the pixels in one frame to another to generate the panoramic frame.

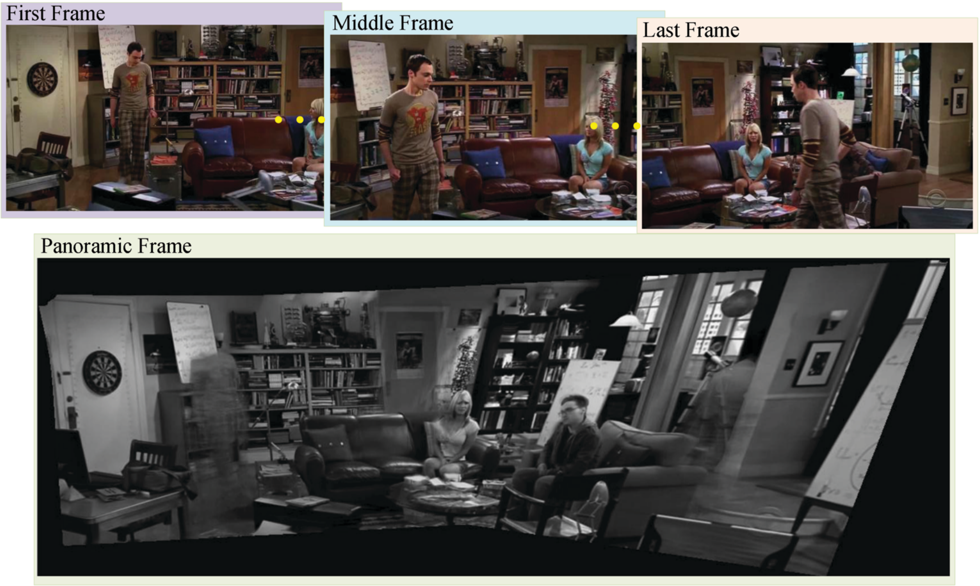


Fig.2. Examples of key-frames using panoramic frames.

More specifically, while $\boldsymbol{F}_t \in \mathbb{R}^{M\times N}$ means the frame at time t and $\boldsymbol{h}_t$ refers to the homography matrix from $\boldsymbol{F}_t$ to $\boldsymbol{F}_{t+1}$. We spatially transfer $\boldsymbol{F}_t$ to $\boldsymbol{F}_{t+1}$ using the operation of $\widetilde{\boldsymbol{F}}_{t+1} = \boldsymbol{F}_t \circ \boldsymbol{h}_t$. Ideally, $\widetilde{\boldsymbol{F}}_{t+1}$ should be approximated to $\boldsymbol{F}_{t+1}$, and the error arising from outliers pixels denoted as $\boldsymbol{e}_t = \widetilde{\boldsymbol{F}}_{t+1} - \boldsymbol{F}_{t+1}$ should be assumed to be sufficiently sparse. So far, the video registration problem is transferred to estimate the optimal sequence of homography matrices that map consecutive frames and render the sparsest error (minimum ℓ_0 norm). Since this problem is NP-hard in general and non-convex especially due to the nonlinear constraints, the cost function is replaced with its convex envelope (ℓ_1 norm) as follows:

$$\begin{aligned} &\min_{\boldsymbol{e}_{t+1}} \|\boldsymbol{e}_{t+1}\|_1 \\ \text{s.t.}:\ &\boldsymbol{F}_t \circ \boldsymbol{h}_t = \boldsymbol{F}_{t+1} + \boldsymbol{e}_{t+1}. \end{aligned} \tag{4}$$

Although the objective function is convex, the equality constraint is still not convex. Thus, with an iteratively solved linearized convex problem, it begins with an estimation of each homography denoted as $\boldsymbol{h_t}^{(k)}$ at the $(k+1)$-th iteration. In order to linearize the constraint around a current estimate of the homography, the current estimation will be $\boldsymbol{h_t}^{(\boldsymbol{k}+1)} = \boldsymbol{h}_{\boldsymbol{t}}^{(\boldsymbol{k})} + \triangle\boldsymbol{h_t}$. Thus, (4) is relaxed to

$$\begin{aligned} &\min_{\triangle\boldsymbol{h_t},\boldsymbol{e}_{t+1}} \|\boldsymbol{e}_{t+1}\|_1 \\ \text{s.t.}:\ &\boldsymbol{J}_t^{(k)}\triangle\boldsymbol{h_t} - \boldsymbol{e}_{t+1} = \boldsymbol{\delta}_{t+1}^{(k)}, \end{aligned} \tag{5}$$

where $\boldsymbol{\delta}_{t+1}^{(k)} = \boldsymbol{F}_{t+1} - \boldsymbol{F}_t \circ \boldsymbol{h}_t^{(k)}$ represents the error incurred at iteration k and $\boldsymbol{J}_t^{(k)} \in \mathbb{R}^{MN\times 8}$ is the Jacobian of $\boldsymbol{F}_t \circ \boldsymbol{h}_t$. Finally, the homography matrix is successfully extracted since the problem in (5) becomes a linear problem which can be solved in polynomial time. After that, the panoramic frame in each shot can be obtained with this estimated homography matrix, as can be seen in Fig.2. More details of the video registration and stitching can be found in [24].

5.2 RFPs Extraction and Representation

Although the panoramic frame has contained more comprehensive contents of a VSC, different panoramic frames may refer to different view directions' appearances. Therefore, this subsection explains the method to extract more representative feature patches (RFPs) from these panoramic frames.

Firstly, movie M is supposed to be segmented into a shot set $\mathcal{T} = \{t_1, \ldots, t_m\}$ and then we cluster shots into the VSC set $\mathcal{S} = \{s_1, \ldots, s_n\}$[20]. Meanwhile, key-frames are obtained using the method depicted in Subsection 5.1 for each shot.

Local features are proved to be more reasonable for semantic analysis and also more robust to sundry variations and occlusions[2]. Thus, we extract two types of local features in key-frames: Scale-Invariant Feature Transform (SIFT) key points and Maximally Stable Extremal Regions (MSER). In the following we describe these two features in detail.

- *SIFT*. We compute the SIFT descriptor for the regions obtained with the DoG detector[25]. We utilize SIFT descriptors with scales of each interest point varying from 20 to 120 pixels. It describes the static appearance over spatial histogram.
- *MSER*. Every extremal region is a connected component of a thresholded image. The regions are obtained by thresholding the intensity image and tracking the connected components as the threshold value changes. The idea is due to the work of Matas *et al.*[26].

The SIFT features are invariant to image scale/rotation and robust to changes in illumination, noise, and minor changes in viewpoint. Meanwhile, SIFT features exhibit the highest matching accuracies for an affine transformation of 50 degrees. But, after this transformation limit, results start to become unreliable. However, MSER as a method of blob detection in images, has been proved as one of the most robust feature detectors on invariance of affine transformation[27]. In order to generate the invariant description for MSER feature, an elliptical image region is used to cover the distinguished regions with the interest point as the center. Meanwhile, the elliptical region normalized to a circle, and the SIFT descriptor for the central point is output as the final MSER feature description. These SIFT and MSER features are named as feature patches.

In addition, human regions always shelter the background, thus the human regions should be considered as the noise or obstacle for movie scene recognition. Therefore, the human detection method[28] is applied to the middle frame of each shot at first, and we exclude shots with a large portion of human region in middle frames, which correspond to the close up shots and the medium close up shots. As shown in Fig.3, for remained shots, we extract the panoramic frames as key-frames, locate the human regions, and then mask them out. Feature patches within the mask are then filtered out, and we reserve feature patches from the remainder. Meanwhile, we cluster the reserved patches in one VSC using K-means algorithm and assign the label of the centroid to each cluster. These labeled patches are named as representative features patches and denoted as $\boldsymbol{R} = \{r_1, \ldots, r_{|\boldsymbol{R}|}\}$, where r_i is a patch, $|.|$ means the size operator, and $i = 1, \ldots, |\boldsymbol{R}|$.

In our approach, the recognition model is based on the Bayesian model using LDA[2]. We use VSCs as the basic data for learning, and each VSC is represented by an RFP set. RFPs are considered as the basic units of codewords, and they are used for training and learning. Besides, the reoccurred VSCs referring to the same location are used to enhance the final recognition results.

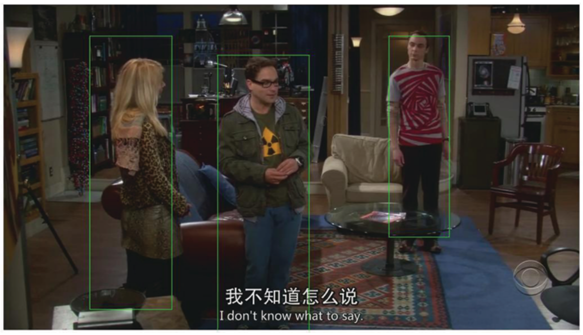


Fig.3. Human detection in key-frames.

6 VSC Correlation Based Recognition

6.1 RFPs Based Individual VSC Recognition

To estimate the LDA topic mixtures, a patch $\boldsymbol{x}$ is the basic feature unit, defined to be a patch membership from a dictionary of codewords indexed by $1, \ldots, T$. The t-th codeword in the dictionary is represented by a T-vector $\boldsymbol{a}$ such that $\boldsymbol{a}^t = 1$ and $\boldsymbol{a}^v = 0$ for $v \neq t$. A patch (or codeword) refers to an RFP. A VSC is a sequence of N patches denoted by $\boldsymbol{x} = (x_1, x_2, \ldots, x_N)$, where x_n is the n-th patch of the VSC. A category is a collection of M VSCs denoted by $\boldsymbol{C} = \{\boldsymbol{a}_1, \boldsymbol{a}_2, \ldots, \boldsymbol{a}_M\}$.

Given an unknown VSC s_i, it is represented by a set of codewords firstly. We empirically find that the quality of topic distributions is relatively stable if the number of topics is within a reasonable range, and we choose $k = 35$ topics for the construction of LDA. Finally, we have the decision probability of s_i to be classified to movie scene category c.

$$p(c|s_i, \boldsymbol{\theta}, \boldsymbol{\beta}, \boldsymbol{\eta}) \propto p(s_i|c, \boldsymbol{\theta}, \boldsymbol{\beta}) p(c|\boldsymbol{\eta}) \propto p(s_i|c, \boldsymbol{\theta}, \boldsymbol{\beta}), \quad (6)$$

where $\boldsymbol{\theta}, \boldsymbol{\beta}$ and $\boldsymbol{\eta}$ are parameters learnt from the training set and c is the category index. Then, we use a decision probability $p(c|s_i)$ to classify VSC s_i into category c. After that, each VSC is classified into a movie scene category with the largest decision probability.

6.2 Enhanced Recognition Based on VSC Correlation

Actually, there are many VSCs referring to the same physical location in a movie. But the features extracted in different VSCs with different appearances may present various recognition ability in the recognition model. However, the similarity correlations between these VSCs are easy to obtain and they are reliable context information for more accurate recognition. Therefore, we use the near duplicate VSC identification to get the VSC correlations. Given two VSCs s_x and s_y, they are represented with two RFP sets $\mathcal{R}_x$ and $\mathcal{R}_y$ respectively. For each RFP in $\mathcal{R}_x$, it involves a series of comparisons to search the nearest neighbor in $\mathcal{R}_y$ and it is computationally expensive. However, the LIP-IS algorithm[29] can be used for more fast RFP nearest neighbor searching. The similarity definition Sim for two RFPs $r_x = \{r_{x,1}, r_{x,2}, \ldots, r_{x,36}\}$ and $r_y = \{r_{y,1}, r_{y,2}, \ldots, r_{y,36}\}$ is defined as

$$Sim(r_x, r_y) = \sum_{i=1}^{36} Col(r_{x,i}, r_{y,i}), \quad (7)$$

where $0 < Col(r_{x,i}, r_{y,i}) \leqslant 1$ is the collision function[29]. The matching score of s_x and s_y is summarized in Algorithm 1.

Algorithm 1. VSC Matching with LIP-IS Filtering Mechanism

```
Input: VSC $s_x$ and $s_y$
Output: matching score of $s_x$ and $s_y$
1: Extract the RFP sets $\mathcal{R}_x$ and $\mathcal{R}_y$ in $s_x$ and $s_y$ respectively.
2: Hash all RFPs in $\mathcal{R}_x$ to LIP-IS[29].
3: for each RFP $r_y$ in $\mathcal{R}_y$ do
4:     Hash $r_y$ to LIP-IS.
5:     Retrieve the nearest neighbor RFP $r_x$ in $\mathcal{R}_x$ and label it as the matched one.
6: Compute the number of matched RFP pairs (i.e., $\{(r_x, r_y)|r_x \in \mathcal{R}_x, r_y \in \mathcal{R}_y\}$).
7: Return the portion of matched RFP pairs as the matching score.
```

With Algorithm 1, we can get the matching score (also named correlation score) $Sc(x, y)$ of s_x and s_y. If $Sc(x, y)$ is bigger than the threshold T (we take $T = 0.68$), they are named as the near duplicate for each other. Supposing the near duplicates s_i and s_t are with decision probability of $p(c|s_i)$ and $p(c|s_t)$ respectively, and their correlation score is $Sc(i, t)$, we update the decision probability of $p(c|s_i)$ as follows.

$$p'(c|s_i) = \frac{1}{N} \sum_{t=1}^{N} \Big(\frac{1}{1 + Sc(i,t)} p(c|s_i) + \frac{Sc(i,t)}{1 + Sc(i,t)} p(c|s_t) \Big), \quad (8)$$

where N is the number of the near duplicate VSCs for s_i, and $p'(c|s_i)$ is the new decision probability for scene s_i referring to the category c.

According to (8), we can see that, if the matching score $Sc(i,t)$ is small, it means that the similarity measure is not reliable enough, and the near duplicate VSC cannot provide credible context information. Otherwise, the two VSCs could be seen as the same location, and their final decision probabilities should be consistent with a weighted average calculation.

7 Experimental Results

For movie scene recognition, the key issue is a sufficient number of movie clips for visual training. Thus, Marszałk *et al.* have provided a dataset with 12 classes of human actions and 10 classes of scenes distributed over 3 669 video clips, using the alignment of scripts and videos. However, this dataset is not suitable for movie scene model training. That is because the video clip is not a complete VSC in this dataset, and the appearance, including illumination, resolution and so on, is not so typical and representative. Thus, we take the idea of using script, to use *scene captions*, short descriptions of the scene setup, to construct our movie scene dataset.

Our dataset contains five categories of movie scene (*Street* (45 clips), *Office* (58 clips), *Restaurant* (46 clips), *Bedroom* (51 clips), *Incar* (61 clips)), because they are the most common ones in movies[7]. This is also shown in Fig.4. In addition, we also use WordNet[30] to select expressions corresponding to the five instances. That means if the scene setting in scripts is *cafe*, it refers to the generalized concept of *restaurant*. In fact, we only use scripts for training and do not assume scripts to be available during learning.

In the following, we mainly introduce how we deal with the five movie scene categories. Each VSC is represented by an RFP set, and each RFP set is named as a movie scene instance for training. While the codebook of codewords is learned, we get the recognition model. While the model is trained, our experiments are performed over 50 movie clips which are generated with our automatic scene segmentation, to assess the performance of our approach. For these 50 clips, each one is a VSC and 10 for each category. The RFPs are extracted in the VSCs and the RFP set is constructed to put into the trained model for recognition.

We first evaluate our approach by comparing it with three others approaches: the movie scene recognition method proposed by Marszałk *et al.*[7] and two image-based scene recognition methods: the spatial pyramid matching based approach[6] and the Centrist-based approach[3], both on Marszałk's dataset (MD) and our scene dataset with five categories (DFC). And also, we perform the comparisons on the following various con-

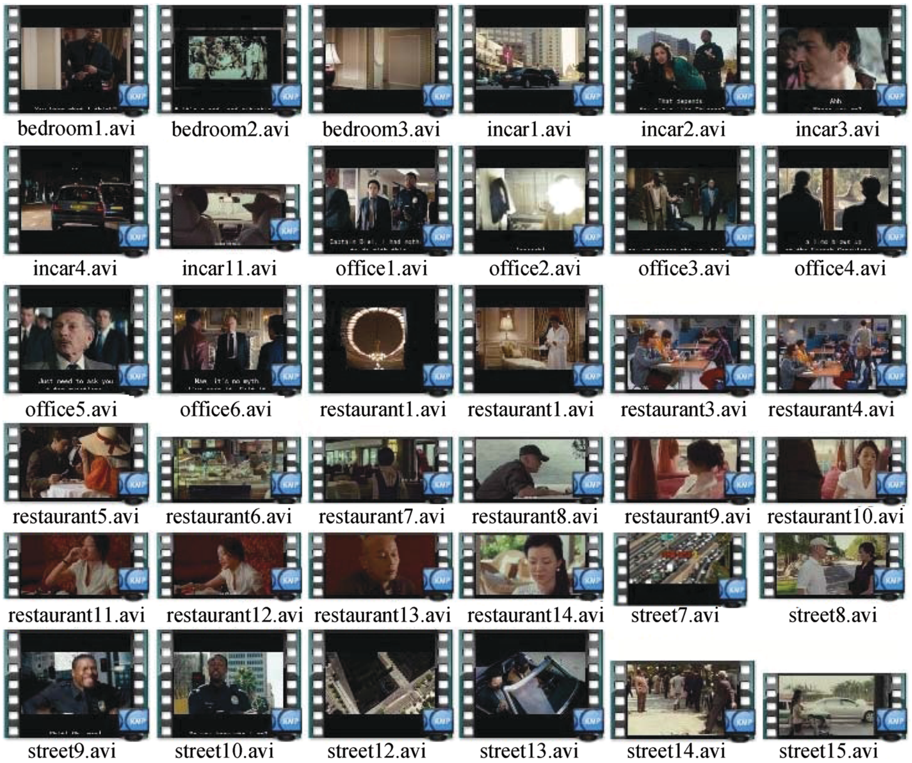


Fig.4. Some example movie scenes in our dataset.

ditions: with and without panoramic frame based key frame extraction, with and without RFP and human region removal, with and without VSC correlation. The first set of comparisons is to assess the overall performance of our approach, and the second set is implemented to prove the efficiency of both panoramic frame based key frame extraction and VSC correlation based recognition improvement.

7.1 Performance of Different Methods

To the best of our knowledge, while most of the scene recognition researches focus on static images, there is a little work directly on movie scene recognition, for example [7]. Thus, in this subsection, we compare the performance of our approach[6] (OUR_M) and the approach in [7] (MARCIN_M). Besides, in order to assess the efficiency of our approach more comprehensively, we also implement the Centrist-based approach[3] (CENTRIST_M) and Spatial Pyramid Matching based approach[6] (SPM_M). Because the last two approaches were proposed to handle image scene recognition, we adopt them to perform video scene recognition by applying them on the mid-frame based key frames in each shot, and each VSC is classified into one of the five scene categories with the largest decision probability. In addition, we also evaluate these three methods in DFC and a subset of MD consisting of 50 randomly chosen clips. The recognition results are shown in Table 1.

Table 1. Comparison of Proposed Approach with MARCIN_M[7], CENTRIST_M[3], and SPM_M[6]

	OUR_M		MARCIN_M		CENTRIST_M		SPM_M	
	MD	DFC	MD	DFC	MD	DFC	MD	DFC
Street	0.81	0.84	0.52	0.55	0.62	0.66	0.59	0.68
Office	0.74	0.80	0.62	0.63	0.60	0.61	0.58	0.61
Restaurant	0.61	0.67	0.33	0.40	-	-	-	-
Bedroom	0.72	0.71	0.51	0.51	0.69	0.70	0.65	0.69
Incar	0.64	0.64	0.66	0.67	-	-	-	-

7.2 Performance Evaluation on Using Panoramic Frame

Considering the use of panoramic frame, we compare the performance of panoramic frame based key frame extraction with normal key frame extraction. Here, the normal key frame extraction is the key frame extraction method in [20].

Static scene images are always landscape images and contain more comprehensive features of the whole scene, but the focus or attention in the movie scene is mostly the moving objects. Thus there are only a few features about the scene in some shots. However, using panoramic frame based key frames, we can collect more comprehensive scene features in consequent frames taking account that "redundance is also abundance". Finally, as shown in Fig.5, we get a more efficient recognition performance which archives an improvement of 9% accuracy on average in the five scene categories.

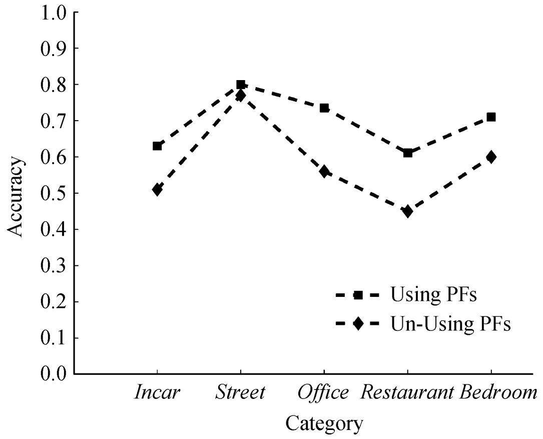
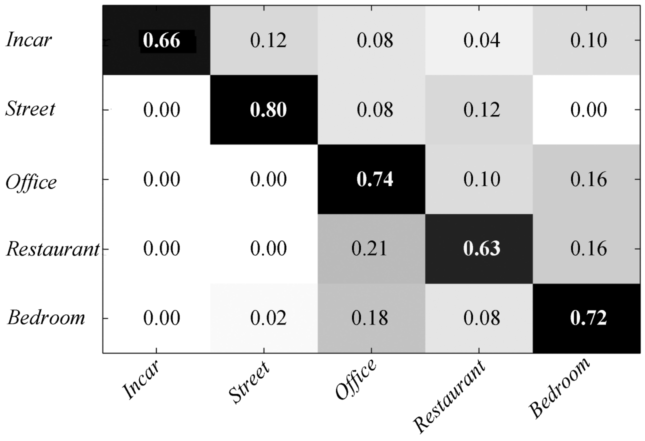


Fig.5. Comparisons of using or not using panoramic frames (PFs).

Besides, the confusion matrix in all categories is also reported in Fig.6. In fact, the most confusing pairs of the five categories are *office/restaurant*, *office/bedroom* and *restaurant/bedroom*, especially for normal key frame extraction. That is because in partial scene, these pairs of categories share very similar scenes or backgrounds. However, from Fig.6, we can see that our approach has satisfactory results in these pairs because different categories present distinguishing appearances in the panoramic frames.



	Incar	*Street*	*Office*	*Restaurant*	*Bedroom*
Incar	0.66	0.12	0.08	0.04	0.10
Street	0.00	0.80	0.08	0.12	0.00
Office	0.00	0.00	0.74	0.10	0.16
Restaurant	0.00	0.00	0.21	0.63	0.16
Bedroom	0.00	0.02	0.18	0.08	0.72

Fig.6. Confusion matrix of our movie scene dataset.

7.3 Comparison on Different Experimental Conditions

In order to assess the effectiveness of the feature extraction on non-human regions as well as the VSC correlation based enhanced recognition, we perform several experiments in different experimental conditions. We evaluate the performance of the RFPs extraction, the human regions removing as well as the VSC correlation based enhanced recognition. The human regions always take a large portion in most of the frames, but they are the noises for scene recognition which mainly refer to background information. Thus, the human region removing processing is introduced to exclude local features in human regions, and the improved recognition accuracy is shown in Fig.7.

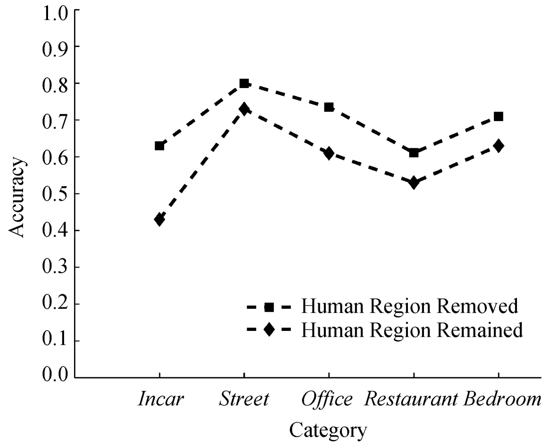


Fig.7. Performance improvement after human region removed.

In Fig.8, we compare the performance of the directly sift features extraction based scene recognition (DSFSR) method, which is the method used in [7], method with our RFP extraction and human region removing (RFEHRSR), and also RFEHRSR combined with the VSC-enhanced recognition method (VERFHRSR). We find that the RFP and the human regions removing enhance the recognition accuracy. Because we reduce most of the redundant features which are identified as noise by RFPs and by removing the human regions, the pure location regions are more efficiently used for training and learning.

Meanwhile, the recognition of outdoor movie scene of *street* even reaches a high accuracy of 80%, while the recognition results of the three indoor movie scenes: *office, restaurant, bedroom*, are not so good as that of *street*. The movie scene of *Incar* seems to be discriminated easily, but the recognition result is not very well. It is maybe because the human region occupies a very large portion in a *Incar* movie scene, and the number of features used for training and recognition is very small. We conclude that without removing the human regions, the performance decreases obviously, especially in inside movie scenes of *bedroom* and *office*. Besides, the VSC correlation enhances the recognition result of individual VSC. Our approach achieves a satisfactory performance on the five movie scene categories as shown in Fig.8.

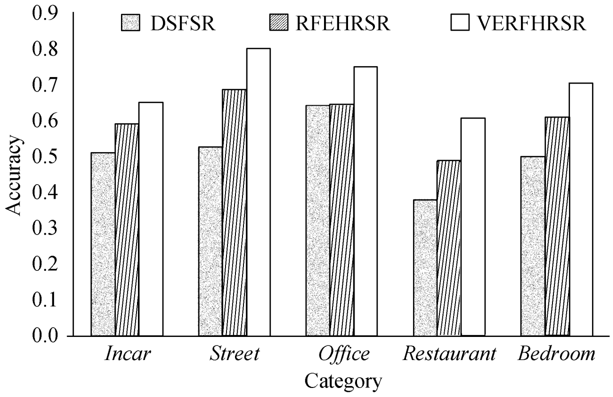


Fig.8. Recognition performances on different conditions.

8 Conclusions

In this paper, we have studied how to effectively recognize scenes in a movie. More specifically, the movie is efficiently segmented into clips at first. Then, by stitching the panoramic frame as key frames, the representative local features in these key-frames are extracted, and further the noise of human regions is removed. In addition, during the process of extracting all the local patches in each VSC, the RFPs are chosen to represent the VSC. After that, an LDA-based movie scene recognition model is built by training the collected VSCs. Finally, when the recognition results for each individual VSC is ready, the correlations of VSCs are taken into account for enhanced recognition. Although the recognition results are not dramatically improved, it is a very meaningful idea to collectively use both key-frame information and related VSC information for video content analysis.

References

[1] Gupta L, Pathangay V, Dyana A, Das S. Indoor versus outdoor scene classification using probabilistic neural network. *EURASIP Journal on Applied Signal Processing*, 2007, 2007(1).

[2] Li F F, Perona P. A Bayesian hierarchical model for learning natural scene categories. In *Proc. the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Jun. 2005, pp.524-531.

[3] Wu J X, Rehg J M. CENTRIST: A visual descriptor for scene categorization. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 2011, 33(8): 1489-1501.

[4] Liu J, Shah M. Scene modeling using co-clustering. In *Proc. the 11th Int. Conf. Computer Vision*, Oct. 2007.

[5] Wu J X, Rehg J M. Where am I: Place instance and category recognition using spatial PACT. In *Proc. the 2008 IEEE Conference on Computer Vision and Pattern Recognition*, Jun. 2008.

[6] Lazebnik S, Schmid C, Ponce J. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *Proc. the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Jun. 2006, pp.2169-2178.

[7] Marszałek M, Laptev I, Schmid C. Actions in context. In *Proc. the 2009 IEEE Conference on Computer Vision and Pattern Recognition*, Jun. 2009, pp.2929-2936.

[8] Engels C, Deschacht K, Becker J H *et al.* Automatic annotation of unique locations from video and text. In *Proc. the 21st British Machine Vision Conference*, Aug. 31 - Sept. 3, 2010.

[9] Zhou X, Zhuang X D, Tang H *et al.* A novel Gaussianized vector representation for natural scene categorization. In *Proc. the 19th Int. Conf. Pattern Recognition*, Dec. 2008, pp.1-4.

[10] Greene M R, Oliva A. Recognition of natural scenes from global properties: Seeing the forest without representing the trees. *Cognitive Psychology*, 2009, 58(2): 137-176.

[11] Xiao J X, Hays J, Ehinger K A *et al.* SUN database: Large-scale scene recognition from abbey to zoo. In *Proc. the 23rd IEEE Conference on Computer Vision and Pattern Recognition*, Jun. 2010, pp.3485-3492.

[12] Ando R, Shinoda K, Mochizuki T. A robust scene recognition system for baseball broadcast using data-driven approach. In *Proc. the 6th ACM Int. Conf. Image and Video Retrieval*, Jul. 2007, pp.186-193.

[13] Huang J C, Liu Z, Wang Y. Joint scene classification and segmentation based on hidden Markov model. *IEEE Trans. Multimedia*, 2005, 7(3): 538-550.

[14] Schaffalitzky F, Zisserman A. Automated location matching in movies. *Computer Vision and Image Understanding*, 2003, 92(2/3): 236-264.

[15] Héritier M, Gagnon L, Foucher S. Places clustering of full-length film key-frames using latent aspect modeling over SIFT matches. *IEEE Trans. Circuits and Systems for Video Technology*, 2009, 19(6): 832-841.

[16] Héritier M, Foucher S, Gagnon L. Key-places detection and clustering in movies using latent aspects. In *Proc. the 14th IEEE Int. Conf. Image Processing*, Sept. 16 - Oct. 19, 2007, pp.225-228.

[17] Bosch A, Zisserman A, Muãoz X. Scene classification via pLSA. In *Proc. the 9th European Conference on Computer Vision*, May 2006, pp.517-530.

[18] Ni K, Kannan A, Criminisi A, Win J. Epitomic location recognition. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 2009, 31(12): 2158-2167.

[19] Gao G Y, Ma H D. Accelerating shot boundary detection by reducing spatial and temporal redundant information. In *Proc. the 2011 IEEE Int. Conf. Multimedia and Expo*, Jul. 2011, pp.1-6.

[20] Gao G Y, Ma H D. Multi-modality movie scene detection using kernel canonical correlation analysis. In *Proc. the 21st Int. Conf. Pattern Recogntion*, Nov. 2012, pp.3074-3077.

[21] Zeng X L, Hu W M, Liy W *et al.* Key-frame extraction using dominant-set clustering. In *Proc. the 2008 IEEE Int. Conf. Multimedia and Expo*, Jun. 2008, pp.1285-1288.

[22] Rasheed Z, Shah M. Detection and representation of scenes in videos. *IEEE Trans. Multimedia*, 2005, 7(6): 1097-1105.

[23] Xiao J X, Ehinger K A, Oliva A, Torralba A. Recognizing scene viewpoint using panoramic place representation. In *Proc. the 2012 IEEE Conf. Computer Vision and Pattern Recognition*, June 2012, pp.2695-2702.

[24] Ghanem B, Zhang T Z, Ahuja N. Robust video registration applied to field-sports video analysis. In *Proc. the 2012 IEEE Int. Conf. Acoustics, Speech and Signal Processing*, March 2012.

[25] Lowe D G. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 2004, 60(2): 91-110.

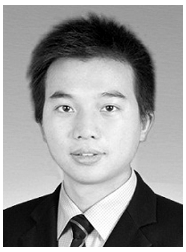
[26] Matas J, Chum O, Urban M, Pajdla T. Robust wide-baseline stereo from maximally stable extremal regions. *Image and Vision Computing*, 2004, 22(10): 761-767.

[27] Mikolajczyk K, Schmid C. Scale & affine invariant interest point detectors. *International Journal of Computer Vision*, 2004, 60(1): 63-86.

[28] Bourdev L, Malik J, Poselets: Body-part detectors trained using 3D pose annotations. In *Proc. the 12th Int. Conf. Computer Vision*, Sept. 29 - Oct. 2, 2009, pp.1365-1372.

[29] Zhao W L, Ngo C, Tan H K, Wu X. Near duplicate keyframe identification with interest point matching and pattern learning. *IEEE Trans. Multimedia*, 2007, 9(5): 1037-1048.

[30] Fellbaum C (editor). Wordnet: An Electronic Lexical Database. MIT Press, 1998.



Guang-Yu Gao is an assistant professor at School of Software, Beijing Institute of Technology, China. He received his Ph.D. degree in computer science and technology from Beijing University of Posts and Telecommunications (BUPT) in 2013, M.S. degree in computer science and technology from Zhengzhou University, China, in 2007. He studied at National University of Singapore as a government-sponsored Joint-Ph.D. student from July 2012 to Apr. 2013. His current research interests include applications of multimedia, computer vision, video analysis, machine learning, and Internet of Things (IoT).

Hua-Dong Ma is a professor and the director of Beijing Key Lab of Intelligent Telecommunications Software and Multimedia, Beijing University of Posts and Telecommunications, China. He received his Ph.D. degree in computer science from Institute of Computing Technology, Chinese Academy of Sciences, Beijing, in 1995, M.S. degree in computer science from Shenyang Institute of Computing Technology, Chinese Academy of Sciences in 1990, and B.S. degree in mathematics from Henan Normal University in 1984. He visited UNU/IIST as a research fellow in 1998 and 1999, respectively. From 1999 to 2000, he held a visiting position in the Department of Electrical Engineering and Computer Science, The University of Michigan, USA. He was a visiting professor at The University of Texas at Arlington from July to September 2004, and a visiting professor at Hong Kong University of Science and Technology from Dec. 2006 to Feb. 2007. His current research focuses on multimedia system and networking, sensor networks and Internet of Things. He has published over 100 papers and 4 books in these fields.