# A Semantic Image Category for Structuring TV Broadcast Video Streams

Jinqiao Wang[1], Lingyu Duan[2], Hanqing Lu[1], and Jesse S. Jin[3]

[1] National Lab of Pattern Recognition
Institute of Automation, Chinese Academy of Sciences, Beijing 100080,China
{jqwang, luhq}@nlpr.ia.ac.cn
[2] Institute for Infocomm Research, 21 Heng Mui Keng Terrace, Singapore 119613
lingyu@i2r.a-star.edu.sg
[3] The School of Design, Communication and Information Technology,
University of Newcastle, NSW 2308, Australia
Jesse.Jin@newcastle.edu.au

**Abstract.** TV broadcast video stream consists of various kinds of programs such as sitcoms, news, sports, commercials, weather, etc. In this paper, we propose a semantic image category, named as Program Oriented Informative Images (POIM), to facilitate the segmentation, indexing and retrieval of different programs. The assumption is that most stations tend to insert lead-in/-out video shots for explicitly introducing the current program and indicating the transitions between consecutive programs within TV streams. Such shots often utilize the overlapping of text, graphics, and storytelling images to create an image sequence of POIM as a visual representation for the current program. With the advance of post-editing effects, POIM is becoming an effective indicator to structure TV streams, and also is a fairly common "prop" in program content production. We have attempted to develop a POIM recognizer involving a set of global/local visual features and supervised/unsupervised learning. Comparison experiments have been carried out. A promising result, F1 = 90.2%, has been achieved on a part of TRECVID 2005 video corpus. The recognition of POIM, together with other audiovisual features, can be used to further determine program boundaries.

## 1 Introduction

The management of large TV broadcast video corpus is a challenging problem. Aiming at effective indexing and retrieval, semantic concepts and ontologies have been proposed to bridge the semantic gap inherent to video content analysis. For example, TRECVID'06 proposed the task of extracting 39 high-level features. In this paper, we focus on the structuring of TV streams. We propose a semantic image category, named as Program Oriented Informative Image Category (POIM), to represent significant visual patterns of images occurring at program boundaries. Referring to Fig. 1 and Fig. 2, POIM can be considered as a visual concept about TV stream structure, an intermediate feature to be combined

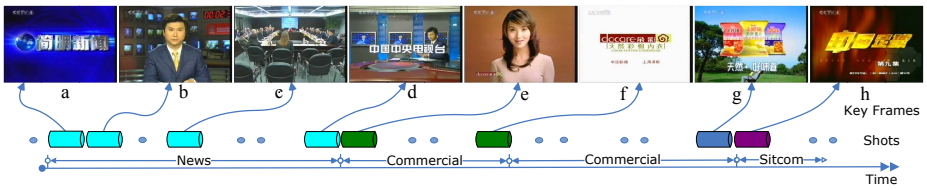


**Fig. 1.** An illustrative example of TV stream structure with POIM images as the anchors of different programs. Shot keyframe a, d, f, g and h are POIM images.

with other audiovisual feature for segmenting different programs and searching for special programs.

In TV broadcast video stream, there are various kinds of programs such as sitcoms, news, sports, commercials, weather, etc. Although different programs are multifarious in video contents, most programs exhibit a prominent structural feature, often in the form of lead-in/-out video shots, at the beginning and/or the end. For example, when we watch a TV program, a sequence of frames marked with the program title can visually indicate the beginning of a program segment after a TV commercial block. On the other hand, a program often comes to the end with frames marked with textual information about its producers and sponsors. Especially in a commercial block, the frames marked with explicit product and company information are often utilized as a prop to highlight what is offered at the end of an individual commercial. Therefore, we generally propose the semantic image category POIM aiming to structure TV streams. In Fig. 2, the functionality of POIM is illustrated by image samples according to different program genres. POIM is a useful concept for distinguishing the transition between different programs in TV streams.


**Fig. 2.** Examples of POIM images from different TV programs (From top to bottom: news, commercial, sitcom, and others).

As illustrated in Fig. 2, POIM is generally classified into four categories: news POIM, commercial POIM, sitcom POIM, and others. News POIM displays the title of news programs, such as "DECISION 200" and "ELECTION NIGHT 200". Commercial POIM provides the brand name, trademark, address, telephone number, and cost, etc. For example, "GEICO" is a brand name and "XEROX" is a company name. An image of a product might be placed with computer graphics techniques. Sitcom POIM shows the name of sitcoms and the producer information. Others POIM show the title of weather report such as "EARTHWATCH" or TV station logos such as "MSNBC" and "LBC".

Let us investigate the example of structuring TV programs from CCTV4 as shown in Fig. 1. From the key frames listed, we note news and sitcom programs start with FMPI shots for providing the names of TV programs. Two commercials end with FMPI shots to catch users' attention by showing product and trademark information. As introduced above, POIM images are dealt with as a useful anchor to represent the structural characteristics of TV streams.

Generally speaking, POIM is composed of text, graphics, and storytelling images. The text is significant, which explicitly provides the information of corresponding TV programs. The graphics and images create a more or less abstract, symbolic, or vivid description of program contents or propagandize a product. The POIM image somehow serves as an introduction poster, by which users can decide whether to continue watching the program or to switch to other TV channels. Moreover, commercial POIM images give a vivid yet brief description of the product and can be considered as the summary of a commercial.

In this paper, we utilize learning algorithms to model the semantic image category POIM on the basis of local and global visual features. As POIM occurs in a sequence of frames, the detection of POIM is applied to key frames within a video shot. For the convenience of description, we call the shot containing at least one POIM image as a POIM shot. Both unsupervised spectral clustering [1] and supervised SVM learning [2] are employed to distinguish POIM images from Non-POIM images. Comparison experiments are carried out.

The rest of this paper is organized as follows. Section 2 presents the extraction of local and global features. Section 3 discusses the unsupervised and supervised learning for training the POIM recognizer. Experiments results are given in Section 4. Finally, Section 5 concludes our work.

## 2 Feature Extraction

The POIM image can be dealt with as a document image involving text, graphics, and storytelling images. We rely on color, edge, and texture features to represent a POIM image. For an incoming TV stream, we perform video shot boundary detection [3]. Loose thresholding is applied to reduce missed video shots. One key frame is simply extracted at the middle of each shot. The extraction of local and global features are carried out within selected key frames.

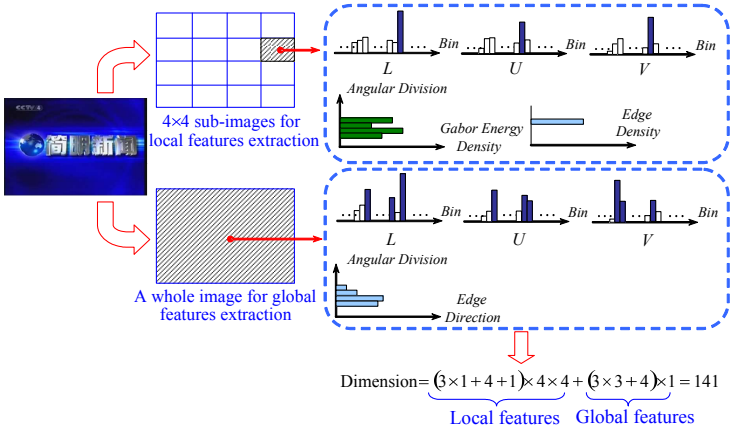


**Fig. 3.** Global feature and local feature extraction for POIM images detection

### 2.1 Global Features

The text, graphics and storytelling image in a POIM image tends to locate at the central part of a frame, as illustrated in Fig. 2. A simple background is often used to highlight the text, graphics, or storytelling image to draw a TV viewer's attention. For global features we take into account the cues of color and edge. CIE LUV color space is used because of its perceptual uniformity. Dominant colors are applied to construct an approximate representation of color distribution. In terms of color we actually utilize a reduced dominant color descriptor. The percentages of selected color bins are taken into account to represent spatial coherency of color. Each channel is uniformly quantized into 100 bins. Three maximum bin values are selected as dominant color features from L, U, and V channels, respectively. This results in a 9-dimensional color feature vector for a whole image. Edge direction histogram is employed to describe the global statistical property of edges. Using Canny edge detection algorithm [4](with $\sigma = 1$, Gaussion mask $Size = 9$), edges are broadly grouped into $h$ categories of orientation by using the angle quantizer as,

$$A_i = \left[ \left\lfloor \frac{180}{h} \right\rfloor \cdot i, \left\lfloor \frac{180}{h} \right\rfloor \cdot (i+1) \right) \qquad i = 0, ...h-1 \tag{1}$$

In our experiment, $h = 4$. Thus 4-dimensional edge direction features are yielded. The total global features are 13 dimensional including color and edge features.

### 2.2 Local Features

As shown in Fig. 2, the spatial distribution is a significant factor in distinguish POIM images. The POIM images first are divided into $4 \times 4$ sub-images, then color, edge and texture features are extracted within each sub-image. The maximum bin value of each channel in the LUV color space is selected as the local

dominant color feature. Note that the bin values are meant to represent the spatial coherency of color, irrespective of concrete color values. The edge density feature for each sub-image is calculated as,

$$Edgedensity_i = \begin{cases} \frac{2E_i}{N} & \text{if } \frac{E_i}{N} \leq 0.5 \\ 1 & \text{else} \end{cases} \tag{2}$$

where $i = 1, \ldots, 16$ is the number of sub-image. $E_i$ is the number of canny edge pixels for sub-image $i$. $N$ is the total number of pixels in sub-image $i$. The local edge density feature are 16 dimensional.

Textures are replications, symmetries and combinations of various basic patterns or local functions, usually with some random variation. Gabor filters [5] exhibit optimal location properties in the spatial domain as well as in the frequency domain, they are used to capture the rich texture in POIM detection. Filtering an image $I(x, y)$ with Gabor filters designed according to [5] results in its Gabor wavelet transform:

$$W_{mn}(x, y) = \int I(x, y) g^*_{mn}(x - x_1, y - y_1) dx_1 dy_1 \tag{3}$$

where $g_{mn}$ is the gabor filters and $*$ indicates the complex conjugate. Within each sub-image, the mean $\mu_{mn}$ of the magnitude of the transform coefficients is used. One scale and four orientations are selected to represent the spatial homogeneousness for each local region. Different from image retrieval and recognition, our task is to find the mutual property of POIM images. One scale and four orientation can reach more satisfactory results than that by more scales and orientations. The total local texture feature are $4 \times 16 = 64$ dimensional.

A 128-dimensional feature vector involving color, edge, and texture is finally formed to represent local features. With the 13-dimensional global feature (9-dimensional dominant color and 4-dimensional edge direction features), we finally construct a 141-dimensional visual feature vector including 128-dimensional local features.

## 3 Detection of POIM Images

The classification of POIM images can be accomplished in both supervised and unsupervised manners. For unsupervised learning, spectral clustering [1] is used to cluster the keyframes into POIM images and non-POIM images. For supervised learning, determinative SVM is employed.

### 3.1 Unsupervised approach

Spectral clustering [1] is a data clustering algorithm developed in recent years. Compared with prior clustering approaches, spectral clustering is a piecewise distance based method and does not assume the data in each cluster having a convex distribution. Also spectral clustering is free of the singularity problem

caused by high dimensionality. The data points are represented as vertices in a graph, and spectral clustering aims at partitioning a graph into disjointed sub-graphs based on the eigenvectors of the similarity matrix. The algorithm is briefly described in Algorithm 1. The affinity matrix $A$ is directly built from the 141-dimensional global and local features. The features are projected into a 2-dimensional space generated by the two largest eigenvectors.

**Algorithm 1 Spectral Clustering**

1. Form the affinity matrix $A \in R^{n \times n}$, $A_{i,i} = 0$, $A_{i,j} = \exp(-||s_i - s_j||^2/2\sigma^2)$.
2. Construct the Laplacian matrix $L = D^{-1/2}AD^{-1/2}$, with degree matrix $D_{i,i} = \sum_j A_{i,j}$.
3. Form the matrix $X = [x_1, x_2, ..., x_k]$ by staking the $k$ largest eigenvectors of $L$.
4. Normalize each row of $X$ to unit length.
5. Cluster each row $X$ into $k$ clusters via K-mean.

### 3.2 Supervised approach

The recognition of a POIM frame is here formulated as a binary classification problem. For supervised learning, SVMs provide good generalization performance and can achieve excellent results on pattern classifications problems. Through a kernel, the training data are implicitly mapped from a low-level visual feature space to a kernel space, and an optimal separating hyperplane is determined therein. This mapping is often nonlinear, and the dimensionality of the kernel space can be very high or even infinite. The nonlinearity and the high dimensionality help SVMs achieving excellent classification performance, especially for linearly nonseparable patterns. Let $\mathbb{R}$ denote the n-dimensional feature space. The training data are $(x_i, y_i), i = 1, \ldots, l$ where $x_i \in \mathbb{R}^n$, $y \in \{-1, +1\}^l$. SVM finds an optimal separating hyperplane that classifies the two classes by the minimum expected test error. The hyperplane has the following form: $\langle w, x_i \rangle + b = 0$, where $w$ and $b$ is the normal vector and bias, respectively. The training feature vectors $x_i$ are mapped into a higher dimensional space by the function $\phi$. The problem of finding the optimal hyperplane is a quadratic programming problem of the following form [2]:

$$\min_{w,\xi} \quad \frac{1}{2} w^T w + C \sum_{i=1}^{n} \xi_i \tag{4}$$

with the constraints $y_i(\langle w, x_i \rangle + b) \geq 1 - \xi_i, \quad i = 1, 2, \ldots, l$. $C$ is the cost controlling the trade-off between function complexity and training error, and $\xi_i$ is the slack variable. The commonly used Gaussian radical basis function (RBF) kernel is optimized to incorporate the prior knowledge of small sample. Gaussian RBF kernel is defined as $exp(-\gamma \|x_i - x_j\|^2), \gamma > 0$. RBF kernel is used in our SVM learning.

## 4 Experiments

To evaluate the performance of POIM recognition, we performed experiments on TRECVID 2005 video database, which is an open benchmark data for video retrieval evaluation. Our training and testing data are taken from 6 general sources: CNN, NBC, MSNBC, NTDTV, LBC and CCTV4. The genres of TV programs include: news, commercial, movie, sitcom, animation, MTV and weather report.

Ground truth of POIM frames was manually labeled. F1 is used to evaluate our algorithm. F1 is defined as follow: $F1 = \frac{2 \times Recall \times Precision}{Recall + Precision}$.

Our POIM image recognizer has achieved an accuracy of $F1 = 81.5\%$ for the unsupervised method, and $F_1 = 90.2\%$ for the supervised method by using 6000 frames comprising 2000 POIM frames and 4000 Non-POIM frames selected from the video database described above. The results of unsupervised spectral clustering are shown in Fig. 4. From Fig. 4, the distribution of POIM images is clearly different from that of Non-POIM images. Moreover, the clustering results shows the effectiveness of our features. For supervised SVMs, this accuracy is

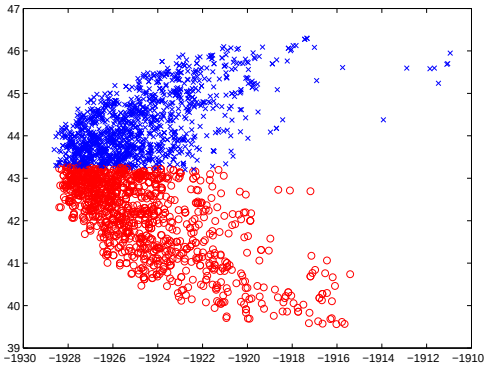


**Fig. 4.** Clustering results of POIM images and non-POIM images. The blue points represent Non-POIM images and the red points represent the POIM images

calculated by averaging the results of ten different random half-and-half training and testing partitions. Radial basis function (RBF) is used for SVMs learning. We are required to tune four parameters, i.e. gamma $\gamma$, cost $C$, class weight $\omega_i$, and tolerance $e$. Class weight $\omega_i$ is for weighted SVM to deal with unbalanced data, which sets the cost $C$ of class $i$ to $\omega_i \times C$. Class weights $\omega_1 = 2$ are set as for POIM and $\omega_0 = 1$ for Non-POIM, respectively. $e$ is set to 0.0001. $\gamma$ is tuned between 0.1 and 10 while $C$ is tuned between 0.1 and 1. As indicated in Fig. 5, "Color" and "Texture" have demonstrated individual capabilities of our color and texture features to distinguish POIM from Non-POIM. Comparatively, texture features play a more important role. The combination of color and texture features results in a significant improvement of performance. An optimal pair $(\gamma, C)$ is selected as $(0.5, 1.0)$ to get an accuracy $F1 = 90.2\%$. By fusing

with audio and textual features, POIM can be further utilized to determine the program boundaries.

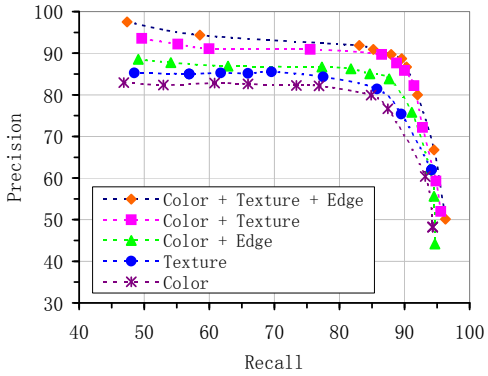


**Fig. 5.** POIM classification performance yielded by using different visual features and different pairs of SVMs parameters $(\gamma, C)$.

## 5 Conclusion

We have proposed a useful image category POIM to facilitate structural analysis in TV streams. An unsupervised approach and a supervised one have been compared for recognizing POIM. A satisfactory accuracy F1=90.2% has been achieved. Future work includes the combination of POIM with audio scene change detection and text content change detection for locating boundaries of TV programs. Moreover, we will explore more effective low-level visual features to improve the recognition performance of POIM.

## 6 Acknowledgement

This work is supported by National Natural Science Foundation of China (Grant No. 60475010 and 60121302).

## References

1. Ng, A.Y., Jordan, M.I., Weiss, Y.: On spectral clustering: analysis and an algorithm. Advance in Neural Information Processing Systems (2001)
2. Vapnik, V.: The nature of statistical learning theory. Springer-Verlag (1995)
3. Wang, J., Duan, L., Lu, H., Jin, J.S., Xu, C.: A mid-level scene change representation via audiovisual alignment. In: Proc. ICASSP'06. (2006)
4. Canny, J.: A computational approach to edge detection. IEEE Trans. PAMI **8**(6) (1986) 679–698
5. Manjunath, B., Ma, W.: Texture features for browsing and retrieval of image data. IEEE Trans. PAMI **18**(8) (1996) 837–842