Support-vector-machine tree-based domain knowledge learning toward automated sports video classification

Guoqiang Xiao
Southwest University
College of Computer and Information Science
Chongqing, Beibei, 400715, China

Yang Jiang
University of Bradford
Digital Media and Systems Research Institute
Richmond Road
Bradford, BD7 1DP, United Kingdom

Gang Song
Southwest University
College of Computer and Information Science
Chongqing, Beibei, 400715, China

Jianmin Jiang
Southwest University
College of Computer and Information Science
Chongqing, Beibei, 400715, China
and
University of Bradford
Digital Media and Systems Research Institute
Richmond Road
Bradford, BD7 1DP, United Kingdom
E-mail: j.jiang1@bradford.ac.uk

Abstract. We propose a support-vector-machine (SVM) tree to hierarchically learn from domain knowledge represented by low-level features toward automatic classification of sports videos. The proposed SVM tree adopts a binary tree structure to exploit the nature of SVM's binary classification, where each internal node is a single SVM learning unit, and each external node represents the classified output type. Such a SVM tree presents a number of advantages, which include: 1. low computing cost; 2. integrated learning and classification while preserving individual SVM's learning strength; and 3. flexibility in both structure and learning modules, where different numbers of nodes and features can be added to address specific learning requirements, and various learning models can be added as individual nodes, such as neural networks, AdaBoost, hidden Markov models, dynamic Bayesian networks, etc. Experiments support that the proposed SVM tree achieves good performances in sports video classifications. © *2010 Society of Photo-Optical Instrumentation Engineers.* [DOI: 10.1117/1.3518080]

Subject terms: machine learning; sports video classifications; support vector machine.

Paper 100042RR received Jan. 25, 2010; revised manuscript received Oct. 11, 2010; accepted for publication Oct. 12, 2010; published online Dec. 7, 2010.

1 Introduction

Sports videos represent an important information source for many entertainment facilities, including TV broadcasting and Internet-based video program consumptions. Automated sports video classification provides important technical tools for a range of applications, such as video indexing, browsing, annotation, and retrieval as well as improvement on their efficiency and effectiveness in accessing sports video archives. At present, extensive research has been carried out and reported on video content analysis and various event detections.[1–12] Reference 1 reported a sports video classification technique via exploitation of the human vision system in perceiving some salient regions inside video frames, which are represented by regions of interests (ROI). The technique first extracts ROIs and then clusters these ROIs to extract color and texture features to classify sports videos. Apart from a complicated algorithm design, the technique does not produce effective classification results observed from the experimental results reported. In Ref. 2, Ma and Zhang reported a motion-based video shot classification technique, which achieved effective results in classifying some motion modes such as jump, run, and other general camera motions. The reported technique can classify video shots in terms of low-level features, but is not able to tell whether such motion modes belong to specific type of sports. In Ref. 3, Geetha and Palanivel described a block-based intensity comparison code for video classification based on a hidden Markov model (HMM). While the technique has the advantage of being robust to illumination changes, the classification is limited to a small number of video patterns. In Ref. 4, supervised learning is reported to build up a semantics dictionary via a K-means clustering technique to classify sports videos, and such a technique could be very complicated, as the size of semantics dictionaries is constantly increasing. A major focus of the existing research is to extract semantics around human objects from sports videos and attempt to add annotations to human object-related videos via classifications.[5,6,8–12] As a result, human object segmentation plays crucial roles in all of these techniques, yet such segmentation itself is a difficult research problem, and existing work is primarily relying on low-level features and detections of their consistency within regions to complete segmentation.[13] Consequently, the accuracy of human object segmentation is limited.

In this work, we propose a SVM tree to illustrate hierarchical and integrated learning through domain-specific knowledge for sports video classification, providing a useful framework for intelligent multimedia content processing, where each individual node has the facility of machine learning relatively independent from other nodes, yet by working

0091-3286/2010/$25.00 © 2010 SPIE

together, can present integrated capture of the overall content features for sports videos. The essential advantage of the proposed approach lies in the fact that the SVM tree has flexibility in terms of both structuring learning nodes and extracting content features, presenting significant potential for tailored needs in multimedia content interpretation and analysis.

The rest of the work is structured into three sections, where Sec. 2 describes the proposed SVM tree design, Sec. 3 reports the experimental results, and Sec. 4 draws conclusions.

2 Proposed Support-Vector-Machine Tree Design

In any video content analysis algorithm development, exploitation of domain knowledge specific to video types remains to an important approach. As an example, surveillance video captured from monitoring a kitchen provides important domain-specific knowledge for the video content analysis, such as "using the washing machine," "food cooking," "washing-up," etc. In the case of sports video classification, individual types of sports present many useful features and provide a range of domain-specific knowledge for exploitation in their content analysis and classifications. Observation of sports videos reveals that spatial distribution of color, texture, and illumination within frames follow certain levels of common rules and common features, which can be highlighted as follows. 1. Dominant color is normally present in the region of audience and the region of sports grounds, and the sports grounds tend to be located in the center of the frames, yet the region of audience tends to be located toward the boundary of frames. 2. The proportion of sports grounds with respect to the region of athletes is different depending on the specific type of sports video, such as football, basketball, etc. 3. The strength of motion presents significant differentiating features among different types of videos. 4. Illumination in the region of audience tends to be weaker in comparison with the illumination in the region of sporting grounds.

To exploit all the domain knowledge as observed and highlighted previously, we propose to divide video frames into blocks and extract features to characterize both the block content and relationships among neighboring blocks to activate machine learning modules for capturing the domain knowledge, and hence achieve learning-based sports video classifications. As the video content presents both spatial and temporal information, and activities described by the video sequences tend to be complicated, where same scenes could be repeated in random occasions, single modules of machine learning, such as neural networks, SVM, etc., would not be able to provide sufficient learning and classification at a global level. To this end, we propose a SVM tree, where an individual SVM module is taken as an internal node inside the tree, and hence its learning is focused on one set of features to characterize or exploit one aspect of the domain knowledge. An overview of our proposed algorithm is illustrated in Fig. 1, and the proposed SVM tree is illustrated in Fig. 2.

From Fig. 1, it can be seen that essentially we propose to use a set of features to represent the domain knowledge, and another set of training videos to optimize the SVM tree structure and enable the tree to learn from the domain knowledge and complete the sports video classification. Such a principle can also be applicable to other types of videos, where domain-specific knowledge is presented, such as news,

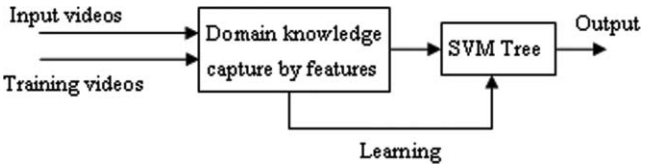

Fig. 1 Illustration of the proposed algorithm.

documentation, etc., which are seen from any broadcasted TV program.

Figure 2 essentially illustrates an example of our initial design to prove the concept and the idea introduced, which can be changed in many different ways, such as adding more learning nodes to accommodate more video types or changing the tree structures, etc. The principle here is to provide hierarchical and integrated learning from the domain knowledge captured by descriptive features, and exploit its localized optimization inside individual SVM nodes, toward optimized learning at a global level via a relatively small set of training videos.

SVM is basically designed for binary classifications.[14–16] Although a number of multiclassification methods have been reported in the published literature, the introduced mechanism of converting binary classification into multiclassification remains limited in terms of full exploitation of SVM power in learning for binary classes. In addition, the complexity of the existing multiple SVM classifiers is normally $O(n)$, yet the proposed SVM tree is $O(\log_2 n)$, where n stands for the number of classification types. Generally, the proposed SVM tree could be arranged as a standard binary tree or a general tree with multiple siblings, which is dependent on the number of sports video types and their descriptive features. As the tree with multiple siblings requires multiple SVM classifiers to be the nodes, which inevitably increases the complexity of the algorithm design, we adopt the binary tree structure to preserve the advantages of binary classifications. The essential purpose is to exploit the flexibility of a tree structure to integrate individual SVM nodes into a global learning structure, and accommodate the varying nature of the sports video content analysis, processing, and classifications.

As seen from Fig. 2, our initial SVM tree includes five internal nodes representing individual SVM learning units, and six external leaf nodes representing classified sports video types, including football, basketball, volleyball, table-tennis, tennis, and badminton. In principle, the number of internal nodes is determined by considering the number of input sports video types to be classified. In other words, the

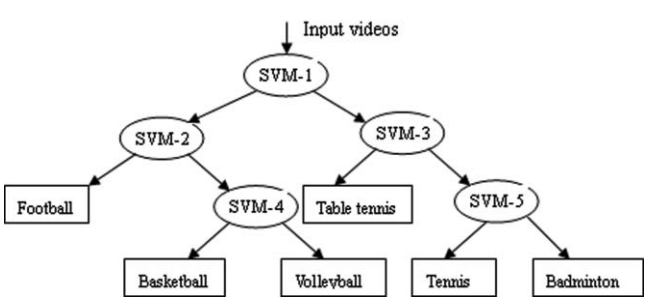

Fig. 2 Illustration of SVM tree.

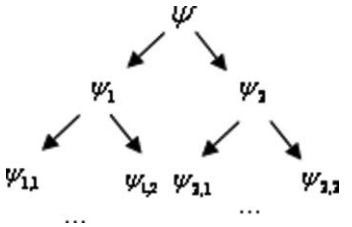

Fig. 3 Illustration of top-down design principle.

number of correspondingly generated leaf nodes should be greater than or equal to the total number of input sports video types, and any number of internal nodes that satisfies the prior condition would be acceptable. In our case, however, this process is manually designed, since the number of sports video types is normally known in advance. Automation of this process can be possible by introducing an automatic inactivation scheme in the training process, where internal nodes can be inactivated and converted into leaf nodes whenever their input training videos only contain one type. As the SVM tree can be extended for classification of more sports types, we used the six types in this work to illustrate the major concept of the reported work.

At the root node represented as SVM-1, the learning procedure should capture the common domain knowledge and classify all the sports video types into two sets: $\psi_1 = \{football, basketball, volleyball\}$, and $\psi_2 = \{table\text{-}tennis, tennis, badminton\}$. At the next level, we have two SVM learning nodes SVM-2 and SVM-3. Depending on what domain knowledge is to be captured by descriptive features and learned by individual SVM nodes, we design SVM-2 as such that the set ψ_1 is classified into one specific type $\psi_{1,1} = football$ and one subset $\psi_{1,2} = \{basketball, volleyball\}$. Similarly, SVM-3 is designed to learn a specific domain knowledge, which classifies the set ψ_2 into one specific sport video type $\psi_{2,1} = table\text{-}tennis$ and one subset $\psi_{2,2} = \{tennis, badminton\}$. Following that, two more internal learning nodes, SVM-4 and SVM-5, are designed to complete the classification of the two subsets $\psi_{1,2}$ and $\psi_{2,2}$.

Given the input sport video set ψ, the proposed SVM tree design follows a top-down principle that video types with similar content or common domain knowledge are grouped together to form two subsets, and each subset is then further grouped into two more subsets according to the common domain knowledge within each individual subset. Such grouping carries on until each specific video type is obtained and hence represented by external or leaf nodes in the SVM tree. The top-down principle is illustrated in Fig. 3.

Following the SVM tree design, the remaining issue is to extract features to represent, describe, and characterize the domain knowledge of each video subset as illustrated in Fig. 3, and drive the SVM node inside the tree for automated learning and classification. As a matter of fact, all machine-learning-based video classification reported in the literature so far share the same principle that features used to characterize the video content are manually selected, although the tree structure is designed by the training process, where a number of possible tree structures are tested and their outcome is used to optimize the tree structure.

Given the fact that numerous features have been reported and researched in the published literature, it becomes difficult to evaluate all these features and select suitable ones for our purposes. Instead of carrying out exhaustive searches and wasting time on the work, we propose a knowledge-based principle to select features for describing domain knowledge observed inside sports videos. The advantage of doing so is to explore the feasibility of the proposed design within manageable categories of features, yet leave sufficient space for future upgrading, and hence making the proposed algorithm flexible and portable enough for further research and investigations. From extensive observations of sports videos, it becomes clear that features of illumination distribution, color distribution, texture, and motion are good starting points for illustrating and implementing the principle of the proposed SVM tree.

To characterize the feature that illumination distribution inside sports videos follows some unique rules where illumination on moving objects tends to be brighter than their surroundings, and human vision systems often rely on the perception of such illumination differences to follow those moving objects, we propose to extract a so-called block intensity comparison code (BICC)[3] inside videos for the purpose of illumination feature extraction. As BICC measures the illumination differences between blocks, it provides good potential for describing and discriminating the illumination distribution among different types of sports videos as well as different video scenes.

Given video frames with the size of $M \times N$, we divide them into $k_1 \times k_2$ blocks, each of which has the size of $h \times v$, where $h = M/k_1$ and $v = N/k_2$. As a result, the value of $k_1 \times k_2$ is the total number of blocks inside each frame, which are determined by the block size. Following the spirit of MPEG, the block size can be selected as 8×8, 16×16, 32×32, and 64×64. While smaller block size enables the capture of regional details, larger block size provides global characterization of correlation between featured regions. In our work, we selected the block size to be 64×64.

The BICC can be generated as a sequence of bits represented by $\eta[j]$ via direct comparison between the mean intensity value of each block and that of every other block inside the video frame, which totals $C^2_{k_1 \times k_2}$ comparisons. Correspondingly, BICC contains $C^2_{k_1 \times k_2}$ bits, and the specific value of each bit is determined as follows:

$$\eta[j] = \begin{cases} 1 & \text{if } \overline{X(m)} > \overline{X(n)}, \quad m \neq n \\ 0 & \text{otherwise} \end{cases}, \tag{1}$$

where $1 \leq m \leq k_1 \times k_2$, $2 \leq n \leq k_1 \times k_2 - 1$, $1 \leq j \leq C^2_{k_1 \times k_2}$; and $\overline{X(m)}, \overline{X(n)}$ are the mean intensity of the compared blocks.

Existing research on color features is extensive for the past few decades, especially in the areas of content-based image retrieval, most of which are based on the principle of recording the statistics of color components via histograms. While some sports video frames may have little difference in terms of illumination distributions, especially those captured in daylight, their color distribution could provide additional cues for differentiation. To reflect the principle of histograms and variation of all the color features reported in the literature, we propose to calculate both the block-based color histogram H_m and the color moments E_m, σ_m, and S_m as the color features for our proposed SVM tree-based classification.

Let $x_{i,c,m}$ be the intensity values of R, G, B components for the i'th pixel inside the m'th block, where $i \in [1, h \times v]$,

$c \in [R, G, B]$, and $m \in [1, k_1 \times k_2]$, then the block-based color histogram can be calculated as:

$$H_{m,c} = [\delta_{1,c}, \delta_{2,c}, \ldots \delta_{j,c} \ldots \delta_{\xi,c}], \tag{2}$$

where $\xi = 32$ stands for the number of quantized intensity categories C_j ($j \in [1, \xi]$), and δ_j counts the number of times that $x_{i,c,m} \in C_j$.

In addition, the second and third order moments from the color components are also calculated to enhance the color description, which are defined as follows:

$$\sigma_{m,c} = \left[\frac{1}{h \times v} \sum_{i=i}^{h \times v} (x_{i,c,m} - \mu_{m,c})^2 \right]^{\frac{1}{2}}, \tag{3}$$

where $\mu_{m,c} = 1/h \times v \sum_{i=1}^{h \times v} x_{i,c,m}$ is the mean value of all pixels inside the m'th block, and $h \times v = 64 \times 64$ defines the block size.

$$S_{m,c} = \left[\frac{1}{h \times v} \sum_{i=i}^{h \times v} (x_{i,c,m} - \mu_{m,c})^3 \right]^{\frac{1}{3}}. \tag{4}$$

For texture description and representation, research is also extensive and widely reported in the literature. For the convenience of minimizing the algorithm complexity, we use moment-based texture description via a co-occurrence matrix[12] extracted from input video frames, details of which are provided as follows.

Let $\beta_c(i, j)_{d,\theta}$ count the number of times that gray-level i occurs with gray-level j at a relative position, where the Euclidean distance between pixels i and j is d at an angle θ, where $i \in [0, 255]$, $j \in [0, 255]$, and $c \in [R, G, B]$. One co-occurrence matrix can be produced to characterize the texture feature if we normalize the count $\beta_c(i, j)_{d,\theta}$, where the values of d and θ are selected empirically through the training process to reflect the texture orientation inside the videos to be classified. Since there are three components inside the color image, the final co-occurrence matrix we produce to describe the texture feature is the weighted summation of all three co-occurrence matrices as shown next:

$$[A] = \left[P(i, j) = \sum_{c=1}^{3} w_c \beta_{i,j}^c \right], \quad \sum w_c = 1, \tag{5}$$

where $P(i, j)$ represents the element of the final co-occurrence matrix, and w_c represents a set of weighting coefficients determined by the transformation between the two color spaces (R, G, B) and (Y, U, V).

As a result, a number of texture features can be extracted from the co-occurrence matrix. First, the second order moment can be extracted to reflect the energy embedded inside the co-occurrence matrix, which measures the distribution of gray levels inside the video frame and the strength of its corresponding texture. The definition is given as:

$$W_1 = \sum_{i=0}^{255} \sum_{j=0}^{255} p^2(i, j). \tag{6}$$

Second, contrast is useful in measuring the appearance of the texture, which can be defined as:

$$W_2 = \sum_{i=0}^{255} \sum_{j=0}^{255} [(i - j)^2 p(i, j)]. \tag{7}$$

To reflect the information embedded inside the texture, we use the concept of entropy to introduce a quantified measurement of the texture distribution, which can be calculated via:

$$W_3 = -\sum_{p(i,j)} P(i, j) \times \log P(i, j). \tag{8}$$

Finally, we could also measure the variances of the texture via the co-occurrence matrix to reflect the level of texture variations among local regions, which often include the normal and inverse variances. Their definitions are given as:

$$W_4 = \sum_{p(i,j)} [P(i, j) - \mu]^2 P(i, j), \tag{9}$$

$$W_5 = \sum_{i=0}^{255} \sum_{j=0}^{255} \frac{P(i, j)}{1 + (i - j)^2}, \tag{10}$$

where μ stands for the mean value.

As discussed earlier, moving objects often dominate the scenes of visual content inside sports videos, providing essential domain knowledge for content analysis and understanding. Although the moving objects are physically occurring in a 3-D world, their projection onto the 2-D space inside all videos still provides important differentiating power for sports video classification, and thus the perspective effects incurred through the projection from 3-D to 2-D are normally small and trivial when measured in terms of the classification results. In practice, such moving objects are primarily represented by moving human objects such as athletes, and relevant targets including football, basketball, etc. To extract features about the moving objects, existing efforts tend to have object segmentation first, followed by a series of other processing modules, such as tracking and detection, etc., which are too complicated for our purpose of learning and classification. As machine learning has already complicated procedures in capturing the properties of the learning targets, any complicated feature extraction process introduced could overload the machine learning modules and make it difficult for practical applications. To this end, we propose a simple motion feature extraction technique via exploiting MPEG motion estimation and compensation techniques for the SVM tree. As the macroblock-based motion vectors reflect the principle of object segmentation, such simple motion features can still reflect or describe the movement nature of the objects among different video scenes and video frames. While direct human observation of such motion features may not provide obvious cues for those moving objects, the machine learning procedure is capable of capturing such moving objects by exploiting the discriminating features among different video scenes and sports types.

Given the motion vector $\rho_{x,y} = (\alpha_x, \alpha_y)$ located at the position of (x, y) inside the video frame, its strength and direction can be described by:

$$\gamma_{x,y} = \sqrt{(\alpha_x)^2 + (\alpha_y)^2}, \tag{11}$$

$$\tan(\theta_{x,y}) = \frac{\alpha_y}{\alpha_x}. \tag{12}$$

Table 1 Assignment of domain knowledge features to SVM nodes.

SVM-1	Motion
SVM-2	Color
SVM-3	Color
SVM-4	Texture ($d = 100, \theta = 0$ deg) and BICC
SVM-5	Texture ($d = 100, \theta = 90$ deg) and BICC

To characterize the strength of global motion inside frames, we quantize $\theta_{x,y}$ into eight directions: $\omega_s = [0, 45, \ldots 315 \text{ deg}], s \in [0 \ldots 7]$; and $\rho_{x,y}$ into eight regions: $\tau_s = [\tau_0, \tau_1, \ldots \tau_7]$. As a result, two motion histograms can be constructed as detailed as:

$$H_\gamma = \frac{1}{k \times k}[\delta_s(\gamma_{x,y} \in \tau_s)], \tag{13}$$

$$H_\theta = \frac{1}{k \times k}[\delta_s(\theta_{x,y} \in \omega_s)], \tag{14}$$

where $\delta_s(\gamma_{x,y} \in \tau_s)$ counts the number of times that $\gamma_s \in \tau_s$, and $s \in [0, 7]$.

By calculating the mean value and variance of the motion vectors, we can obtain two additional statistics features for the motion information inside the video frames. After all the features are extracted, they are assigned to the individual SVM nodes for classification by following the top-down design principle, as shown in Fig. 3. Details of the assignment are given in Table 1.

3 Experimental Results

To evaluate the proposed SVM tree, we implemented the algorithm in a Microsoft VC++6.0 environment under the Windows XP system, and established a video database from archives of TV sports broadcastings. A description of such sports video databases is highlighted in Table 2, where a total of 1592 shots, lasting around 5 to 30 s each, are extracted from all the videos,[17] and the frame size is 352×240.

To measure the performances of the proposed SVM tree, we adopt the widely known precision and recall rates for presenting our experimental results.

To compare with the existing techniques, we choose the work reported in Ref. 1 as our benchmark. As a result, the experimental results achieved by the proposed algorithm are listed in Table 3 and the results achieved by the benchmark are listed in Table 4. From both tables, it can be seen that the proposed algorithm outperforms the existing benchmark, which supports the idea and the principle introduced and described in the work, that when extracted descriptive features are used to drive the SVM tree rather than a single machine learning unit, the integrated learning nodes can provide good performances in sports video classification.

Table 2 Description of sports video shot database.

Video class	Source	Total	Train
Football	Champions League2009, English Premier League, Italian Football League, Spain Soccer League; 2002 World Cup	358	100
Badminton	2008 England Open Badminton; 2008 Asian Badminton Championships	154	82
Tennis	2008, 2009 Australian Open; 2008 Wimbledon	242	100
Basketball	2009 NBA	292	100
Volleyball	2004 Athens Olympics; 2008 Beijing Olympics	169	75
Table-tennis	2004 Athens Olympics; 2008, 2009 World Table Tennis Championships	214	100

4 Conclusions

We describe a SVM tree approach to construct a hierarchical and integrated learning structure to classify sports videos and achieve good performance, as supported by the experimental results. In comparison with existing learning approaches as reported in the literature, the proposed SVM tree has the

Table 3 Experimental results for the proposed SVM tree.

Input videos	Badminton	Basketball	Football	Table-tennis	Tennis	Volleyball	Precision	Recall
Badminton(72)	68				3	1	0.9444	0.9444
Basketball(192)		184	5	3			0.9583	0.9388
Football(258)		5	250			3	0.9690	0.9690
Table-tennis(114)	2			109	3		0.9561	0.8790
Tennis(142)		2		12	128		0.9014	0.9552
Volleyball(94)	2	5	3			84	0.8936	0.9545

Table 4 Experimental results for the benchmark.

Input videos	Badminton	Basketball	Football	Table-tennis	Tennis	Volleyball	Precision	Recall
Badminton(72)	58	1	1	8	4		0.8055	0.8406
Basketball(192)		177	7	2		6	0.9218	0.8806
Football(258)		8	244		2	4	0.9457	0.9242
Table-tennis(114)	2	2		97	13		0.8509	0.8220
Tennis(142)	8	1	1	11	121		0.8521	0.8643
Volleyball(94)	1	12	11			70	0.7447	0.8750

following advantages: 1. low-level complexity in design and computing, i.e., $O(\log_2 n)$ versus $O(n)$; 2. integrated learning yet exploits individual strength; 3. flexibility in structure, where different numbers of nodes and features can be added to address the level of learning, the number of video types to be classified, and the number of features to be extracted; and 4. flexibility in choosing learning modules, where other learning models can be added or replaced such as neural networks, AdaBoost, HMM, and DBN.

Acknowledgment

The authors wish to acknowledge the financial support from European Framework-7 Programme under the project HERMES (contract 216709) and MICIE (contract 225353).

References

1. K. Rapantzikos, N. Tsapatsolis, and Y. Avrithis, "Spatiotemporal saliency for video classification," *Sig. Process. Image Commun.* **24**(7), 557–571 (2009).
2. Y. Ma and H. Zhang, "Motion pattern based video classification and retrieval," *EURASIP J. Appl. Signal Process.* **2**, 199–208 (2003).
3. M. K. Geetha and S. Palanivel, "A novel block intensity comparison code for video classification and retrieval," *Expert Syst. Appl.* **36**, 6415–6420 (2009).
4. R. Cavet, "Revealing the connoted visual code a new approach to video classification," *Computer Graph.* **28**, 361–369 (2004).
5. M. Roh, "Gesture spotting for low-resolution sports video annotation," *Patt. Recog.* **41**, 1124–1137 (2008).
6. X. Gao, "Shot-based video retrieval with optical flow tensor and HMMs," *Patt. Recog. Lett.* **30**, 140–147 (2009).
7. J. Jiang, K. Qiu, and G. Xiao, "An edge block content descriptor for MPEG compressed videos," *IEEE Trans. Circuits, Syst. Video Technol.* **18**(7), 994–998 (2008).
8. L. Y. Duan, M. Xu, Q. Tian, C. S. Xu, and J. S. Jin, "A unified framework for semantic shot classification in sports video," *IEEE Trans. Multimedia* **7**(6), 1066–1083 (2005).
9. H. C. Shih and C. L. Huang, "MSN: statistical understanding of broadcasted baseball video using multi-level semantic network," *IEEE Trans. Broadcast.* **51**(4), 449–459 (2005).
10. C. C. Lien, C. L. Chiang, and C. H. Lee, "Scene-based event detection for baseball videos," *J. Visual Commun. Image Rep.* **18**(1), 1–14 (2007).
11. E. Kijak, G. Gravier, L Oisel, and P. Gros, "Audiovisual integration for tennis broadcast structuring," *Multimedia Tools Appl.* **30**(3), 289–311 (2006).
12. J. C. Felipe, "Retrieval by content of medical images using texture for tissue identification," *Computer-Based Med. Syst.* **16**, 175–180 (2003).
13. A. B. Chan and N. Vasconcelos, "Modeling, clustering and segmenting video with mixtures of dynamic textures," *IEEE Trans. Patt. Anal. Mach. Intell.* **30**(5), 909–926 (2008).
14. Y. Jiang, J. Jiang, and I. Palmer, "Computerized Interactive Gaming via Supporting Vector Machines," *Intl. J. Computer Games Technol.* **4**, 186941 (2008).
15. Y. Liu, "A novel and quick SVM-based multi-class classifier," *Patt. Recog.* **39**, 2258–2264 (2006).
16. C. J. C. Burges, "A tutorial on support vector machines for pattern recognition," *Data Mining Knowledge Discovery* **2**(2), 121–167 (1998).
17. J. Ren and J. Jiang, "Hierarchical modelling and adaptive clustering for real-time summarization of rush videos," *IEEE Trans. Multimedia* **11**(5), 906–917 (2009).

Guoqiang Xiao received the PhD degree in signal and information processing from the University of Electronic Science and Technology of China, Chengdu, and BE degree in radio technology from Chongqing University, Chongqing, China, in 1999 and 1986, respectively. Since 1986, he has been with the College of Computer and Information Science, Southwest University, Chongqing, China, where he is currently a professor. From 2001 to 2003, he was with the Department of Electrical and Electronic Engineering, the University of Hong Kong, as a postdoctor. His research interests include image processing, pattern recognition, neural networks, and wireless network communication.

Yang Jiang completed her undergraduate studies during 2002 to 2006 at the Central University of Nationalities in Beijing, China. Since 2007, she has been a PhD research student specializing in digital media at the Digital Media and Systems Research Institute, University of Bradford, United Kingdom. Her research interests include digital video technologies, machine learning, computer animation, and game designs.

Gang Song is a MSc research student at the Faculty of Computing and Information Science, Southwest University, Chongqing, China. His research interests include signal processing, information retrieval, machine learning, and video content analysis.

Jianmin Jiang received the BSc degree from Shandong Mining Institute, China, in 1982, MSc degree from China University of Mining and Technology in 1984, and PhD from the University of Nottingham, United Kingdom, in 1994. Since 2002, he has been with the University of Bradford as a chair professor of digital media and the director of the Digital Media and Systems Research Institute. His research interests include, image/video processing in compressed domain, video content analysis, stereo imaging, medical imaging, machine learning, and AI applications in digital media processing, retrieval, and analysis. He has published more than 300 refereed research papers and invented one world-wide filed by British Telecom Research Laboratory.