Content Modelling for Human Action Detection via Multidimensional Approach

Lili N. A. liyana@fsktm.upm.edu.my
Department of Multimedia
Faculty of Computer Science & Information Technology
University Putra Malaysia
43400 UPM Serdang
Selangor, Malaysia

Fatimah K. fatimahk@fsktm.upm.edu.my
Department of Multimedia
Faculty of Computer Science & Information Technology
University Putra Malaysia
43400 UPM Serdang
Selangor, Malaysia

Abstract

Video content analysis is an active research domain due to the availability and the increment of audiovisual data in the digital format. There is a need to automatically extracting video content for efficient access, understanding, browsing and retrieval of videos. To obtain the information that is of interest and to provide better entertainment, tools are needed to help users extract relevant content and to effectively navigate through the large amount of available video information. Existing methods do not seem to attempt to model and estimate the semantic content of the video. Detecting and interpreting human presence, actions and activities is one of the most valuable functions in this proposed framework. The general objectives of this research are to analyze and process the audio-video streams to a robust audiovisual action recognition system by integrating, structuring and accessing multimodal information via multidimensional retrieval and extraction model. The proposed technique characterizes the action scenes by integrating cues obtained from both the audio and video tracks. Information is combined based on visual features (motion, edge, and visual characteristics of objects), audio features and video for recognizing action. This model uses HMM and GMM to provide a framework for fusing these features and to represent the multidimensional structure of the framework. The action-related visual cues are obtained by computing the spatio-temporal dynamic activity from the video shots and by abstracting specific visual events. Simultaneously, the audio features are analyzed by locating and compute several sound effects of action events that embedded in the video. Finally, these audio and visual cues are combined to identify the action scenes. Compared with using single source of either visual or audio track alone, such combined audio-visual information provides more reliable performance and allows us to understand the story content of movies in more detail. To compare the usefulness of the proposed framework, several experiments were conducted and the results were obtained by using visual features only (77.89% for precision;

72.10% for recall), audio features only (62.52% for precision; 48.93% for recall) and combined audiovisual (90.35% for precision; 90.65% for recall).

Keywords: audiovisual, semantic, multidimensional, multimodal, hidden markov model.

1. INTRODUCTION

Video can transfer a large amount of knowledge by providing combination of text, graphics, or even images. Therefore, it is necessary to analyze all types of data: image frames, sound tracks, texts that can be extracted from image frames and spoken words. This usually involves segmenting the document into semantically meaningful units, classifying each unit into a predefined scene type, and indexing and summarizing the document for efficient retrieval and browsing. In this research, recent advances in using audio and visual information jointly for accomplishing the above tasks were reviewed. Audio and visual features that can effectively characterize scene content, present selected algorithms for segmentation and classification, and reviews on some test bed systems for video archiving and retrieval will be described. To date, there is no “perfect” solution for a complete video data-management and semantic detection technology, which can fully capture the content of video and index the video parts according to the contents, so that users can intuitively retrieve specific video segments.

However, without appropriate techniques that can make the video content more accessible, all these data are hardly usable. There are still limited tools and applications to describe, organize, and manage video data. Research in understanding the semantics of multiple media will open up several new applications (Calic et al. 2005). Multimodality of the video data is one of the important research topics for the database community. Videos consist of visual, auditory and textual channels, (Snoek et al. 2005). These channels bring the concept of multimodality. Definition of the multimodality given by Snoek et. al. as the capacity of an author of the video document to express a predefined semantic idea, by combining a layout with a specific content, using at least two information channels. The visual part of the video is used to represent something happening in the video. Most probably, one or more entities in the video are acting. The audio part of the video is used to represent things heard in the video. Most probably, some sound is created by some entities either saying or making some sound. The textual part of the video is used to represent something either written on the screen or on some entity in video. All of these visual, audio and textual modalities should be considered in a multimodal video data model.

Modeling and storing multimodal data of a video is a problem, because users want to query these channels from stored data in a video database system (VDBS) efficiently and effectively. Video databases can be better accessed if the index generated contains semantic concepts. So, semantic-based retrieval becomes a proper solution to handling the video database. However, there is only few approaches use all of the information. When either audio or visual information alone is not sufficient, combining audio and visual clues may resolve the ambiguities in individual modalities and thereby help to obtain more accurate answers. In this research, an idea of integrating the audio features with visual features for video detection and retrieval was presented. For feasible access to this huge amount of data, there is a great need to annotate and organize this data and provide efficient tools for browsing and retrieving contents of interest. An automatic classification of the movies on the basis of their content is an important task. For example, movies containing violence must be put in a separate class, as they are not suitable for children. Similarly, automatic recommendation of movies based on personal preferences will help a person to choose the movie of his interest and leads to greater efficiency for indexing, retrieval, and browsing of the data in the large video archives. Beside visual, audio and textual modalities, video has spatial and temporal aspects. In this research, these aspects are also considered.

Spatial aspect is about position of an entity in a specific frame through the video. Spatial position in a specific frame can be given by two-dimensional coordinates. Temporal aspect is about time of a specific frame through the video. Hence a video element can be identified in a video with its frame position(s), X coordinate in frame(s), Y coordinate in frame(s). Specific events that occur in a certain place during a particular interval of time are called video events. Video events occur in a particular shot of the video. As a result, particularly, every event belongs to directly to some specific shot and indirectly to some specific sequence. But they still suffer from the following challenging problems: semantic gap, semantic video concept modelling, semantic video classification, semantic detection and retrieval, and semantic video database indexing and access.

Semantic context detection is one of the key techniques to facilitate efficient multimedia retrieval (Chu et al. 2004). Semantic context is a scene that completely represents a meaningful information segment to human beings. In this research, a multidimensional semantic detection and retrieval approach that models the statistical characteristics of several audiovisual events, over a time series, to accomplish semantic context detection was proposed. This semantic information can be used to produce indexes or tables-of-contents that enables efficient search and browsing of video content.

Action is the key content of all other contents in the video. Action recognition is a new technology with many potential applications. Recognizing actions from videos is important topic in computer vision with many fundamental applications in video surveillance, video indexing and social sciences (Weindland et al. 2006). Actions are described by a feature vector comprising both trajectory information (position and velocity), and a set of other descriptors (Robertson, N & Reid, I. 2006). During the last few years, several different approaches have been proposed for detection, representation and recognition, and understanding video events (Yilmaz, A. & Shah, M. 2005).

Understanding activities of objects, especially humans, moving in a scene by the use of video is both a challenging scientific problem and a very fertile domain with many promising applications. The important question in action recognition is which features should be used? Use of both audio and visual information to recognize actions of human present might help to extract information that would improve the recognition results. What to argue is that action is the key content of all other contents in the video. Just imagine describing video content effectively without using a verb. A verb is just a description (or expression) of actions. Action recognition will provide new methods to generate video retrieval and categorization in terms of high-level semantics.

When either audio or visual information alone is not sufficient, combining audio and visual features may resolve the ambiguities and to help to obtain more accurate answers. Unlike the traditional methods (Chen et al. 2003; Xiong et al. 2003; Divakaran et al. 2005; Zhang et al. 2006; Lu et al. 2001; Li 2000; Zhang & Kuo 2001) that analyze audio and video data separately, this research presents a method which able to integrate audio and visual information for action scene analysis. A multidimensional layer framework was proposed to detect action scene automatically. The approach is top-down for determining and extract action scenes in video by analyzing both audio and visual data. The first level extracts low level features such as motion, edge and colour to detect video shots and next we use Hidden Markov model (HMM) to detect the action. An audio feature vector consisting of *n* audio features which is computed over each short audio clip for the purpose of audio segmentation was used too. Then it is time to decide the fusion process according to the correspondences between the audio and the video scene boundaries using an HMM-based statistical approach. Results are provided which prove the validity of the approach. The approach consists of two stages: audiovisual event and semantic context detections. HMMs are used to model basic audio events, and event detection is performed. Then semantic context detection is achieved based on Gaussian mixture models, which model the correlations among several action events temporally. It is the interest of this research to investigate, discover new findings and contribute the idea to the domain knowledge. This research is a fundamental research with experimental proving. The experimental evaluations indicate that the approach is

effective in detecting high-level semantics such as action scene. With this framework, the gaps between low-level features and the semantic contexts were bridged.

2. PREVIOUS WORKS

Human tracking and, to a lesser extent, human action recognition have received considerable attention in recent years. Human action recognition has been an active area of research in the vision community since the early 90s. The many approaches that have been developed for the analysis of human motion can be classified into two categories: model-based and appearance-based. A survey of action recognition research by Gavrila, in [5], classifies different approaches into three categories: 2D approaches without shape models, 2D approach with shape models and 3D approaches; the first approach to use 3D constraints on 2D measurements was proposed by Seitz and Dyer in [7].

Many approaches have been proposed for behaviour recognition using various methods including Hidden Markov Model, finite state automata, context-free grammar, etc. [8] made use of Hidden Markov models to recognize the human actions based on low-resolution image intensity patterns in each frame. These patterns were passed to a vector quantizer, and the resulting symbol sequence was recognize using a HMM. Their method did not consider the periodicity information, and they have no systematic method for determining the parameters of the vector quantization. [9] presented a method to use spatio-temporal velocity of pedestrians to classify their interaction patterns. [10] proposed probabilistic finite state automata (FA) for gross-level human interactions.

Previous works on audio and visual content analysis were quite limited and still at a preliminary stage. Current approaches for audiovisual data are mostly focused on visual information such as colour histogram, motion vectors, and key frames [1, 2, 3]. Although such features are quite successful in the video shot segmentation, scene detection based on the visual features alone poses many problems. Visual information alone cannot achieve satisfactory result. However, this problem could be overcome by incorporating the audio data, which may have additional significant information. For example, video scenes of bomb blasting should include the sound of explosion while the visual content may vary a lot from one video sequence to another. The combination of audio and visual information should be of great help to users when retrieving and browsing audiovisual segments of interest from database. Boreczky and Wilcox [4] used colour histogram differences, and cepstral coefficients of audio data together with a hidden Markov model to segment video into regions defined by shots, shot boundaries and camera movement within shots.

3. METHODOLOGY

Unlike previous approaches, our proposed technique characterizes the scenes by integration cues obtained from both the video and audio tracks. These two tracks are highly correlated in any action event. We are sure that using joint audio and visual information can significantly improve the accuracy for action detection over using audio or visual information only. This is because multimodal features can resolve ambiguities that are present in a single modality.

Besides, we modelled them into multidimensional form. The audio is analysed by locating several sound effects of violent events and by classifying the sound embedded in video. Simultaneously, the action visual cues are obtained by computing the spatio-temporal dynamic activity signature and abstracting specific visual events. Finally, these audio and visual cues are combined to identify the violent scenes. The result can be seen in Table 6 on comparison recognition percentage with using single source either visual or audio track alone, or combined audio-visual information.

The framework contains two levels. The first level is to compute features and segmentation. The second level is to apply the computation algorithms to the classification, detection and recognition function to obtain the desired action. See Figure 1.

In particular, this part of elaboration will be concerning with the process of action detection and classification, the framework and structure of building the HMM based framework, and with the automatic recognition and retrieval of semantic action detection, as in Figure 1. Through techniques to be proposed, scenes, segments and individual frames can be characterized in video. This research is concerning on the development of HMM based framework for semantic interpretation of video sequences using:

- Edge feature extraction
- Motion
- Colour feature extraction
- Audio

3.1 The First Layer

Basically the task of the first level is to split the image/video data into several regions based on colour, motion, texture or audio information. The goal of the feature extraction process is to reduce the existing information in the image into a manageable amount of relevant properties. First, what we require is the shot boundary between two different frames is clearly detected. We use a pixel-based approach for conducting the shot detection. Assume X and Y are two frames, and d(X,Y) is the difference between the two frames. $P_X(m,n)$ and $P_Y(m,n)$ represent the values of the (m,n)-th pixels of X and Y, respectively. The following equation shows the approach to do shot change detection. below graphical objects, as in Figure 1.

$$d_{XY}(m,n) = \begin{cases} 1, & |P_X(m,n) - P_Y(m,n)| > T_1 \\ 0, & else \end{cases}$$

and (1)

$$d(X,Y) = \frac{1}{m*n}\sum_m \sum_n d_{XY}(m,n) \quad (2)$$

If d(X,Y) is larger than a threshold T_1, a shot change is detected between frames X and Y.

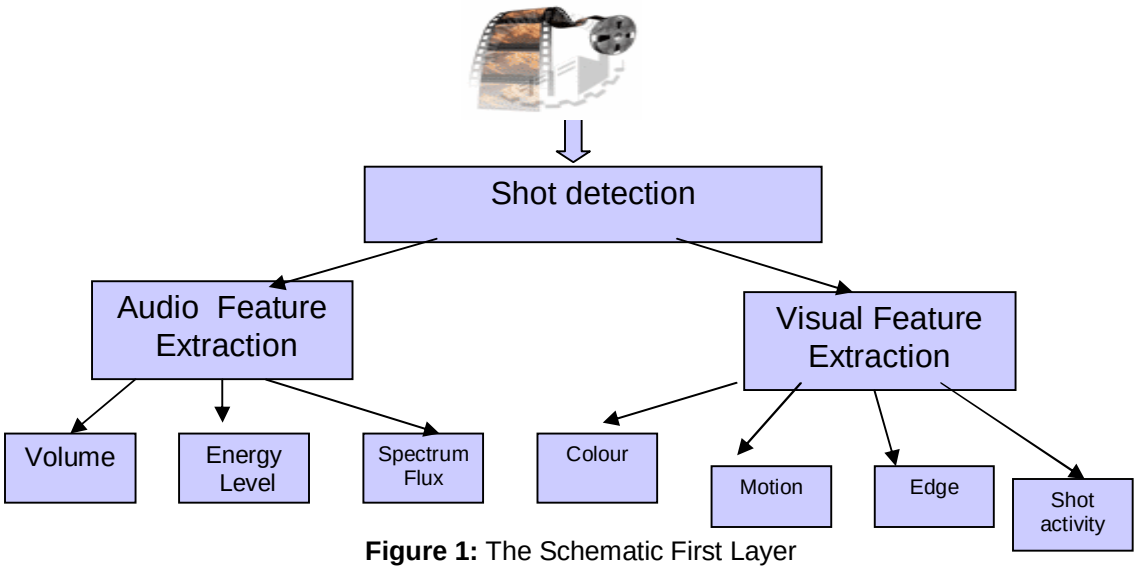

Figure 1: The Schematic First Layer

For extracting audio features, we use an audio feature vector consisting of n audio features which is computed over each short audio clips (100ms duration). The extracted features include: volume, band energy ratio, zero-crossing rate, frequency centroid and bandwidth. These adopted audio features have been widely used in many audio applications, e.g, [2,9] and are known to perform reasonably well.

Volume is the total spectrum power of an audio signal at a given time and is also referred as loudness. It is easy to compute for each frame and is a useful feature to detect silence or to distinguish speech from non-speech signals. The definition of volume is:

$$Vol = \frac{\int_0^{\omega_s} |SF(\omega)|^2 d\omega}{Vol_{max}} \quad (3)$$

Where $SF(\omega)$ denotes the short-time Fourier Transform coefficients and ω_s is the sampling frequency.

The band energy (BE) is defined as the energy content of a signal, in a band of frequencies:

$$BE = \int_{\omega_L}^{\omega_U} |SF(\omega)|^2 d\omega \quad (4)$$

In order to model the characteristics of spectral distribution more accurately, the band energy ratio is considered in the system. The frequency spectrum is divided into sub-bands with equal frequency intervals, then the sub-bands energy are computed and normalized by the frame volume as:

$$BER = \frac{\int_{\omega_L}^{\omega_U} |SF(\omega)|^2 d\omega}{Vol} = \frac{BE}{Vol} \quad (5)$$

where ω_U and ω_L are the upper and the lower bound frequencies of a sub-band, respectively.

Zero crossing rates has been extensively used in many audio processing algorithms, such as voiced and unvoiced components discrimination, end-point detection, audio classification, etc. This feature is defined as the average number of signal sign changes in an audio frame:

$$ZCR = \frac{1}{2N} \sum_{i=1}^{N-1} |sign(x(i)) - sign(x(i-1))| \quad (6)$$

where *x(i)* is the input audio signal, *N* is the number of signal samples in a frame, and *sign()* is the sign function.

Frequency centroid (FC) is the first order statistics of the spectrogram, which represents the power-weighted median frequency of the spectrum in a frame. It is formulated as follows:

$$FC = \frac{\int_0^{\omega_c} \omega |SF(\omega)|^2 \, d\omega}{\int_0^{\omega_c} |SF(\omega)|^2 \, d\omega} \quad (7)$$

Bandwidth (BW) is the second-order statistics of the spectrogram, which represents the power-weighted standard deviation of the spectrum in a frame. The definition of BW is as follows:

$$BW = \sqrt{\frac{\int_0^{\omega_c} (\omega - FC)^2 \, |SF(\omega)|^2 \, d\omega}{\int_0^{\omega_c} \{SF(\omega)|^2 \, d\omega}} \quad (8)$$

Frequency centroid and bandwidth are usually combined to describe statistical characteristics of the spectrum in a frame, and their reliability and effectiveness have been proved in previous work [6]. The video feature extraction, mainly based on the low level visual features, which characterize colours, shapes, textures or motion. Colour breaks are detected by comparing the colour histograms between adjacent video frames. Colour histogram, which represents the colour distribution in an image, is one of the most widely used colour features. The histogram feature measures the distance between adjacent video frames based on the distribution of luminance levels. It is simple, easy to compute, and works well for most types of video [8].

The luminance of a pixel *Lpixel* is computed from the 8-bit red (R), green (G), and blue (B) components as

$$Lpixel = 0.30008(R) + 0.5859(G) + 0.1133(B) \quad (9)$$

H is a 64 bin histogram computed by counting the number of pixels in each bin of 4 gray levels, thus

$$H[k] = \#\ of\ pixels\ where\ k = Lpixel/4,\ 0<=k<=63 \quad (10)$$

As for the texture, we use three values: coarseness, contrast and direction as defined by Tamura [3] to represent its feature.

Time series motion data of human's whole body is used as input. Every category of target action has a corresponding model (action model), and each action model independently calculates the likelihood that the input data belongs to its category. Then the input motion is classified to the most likely action category. The feature extraction (position, direction, movement) focuses attention on the typical motion features of the action, and a model of the features' behaviour in the form of HMM.

A motion data is interpreted at various levels of abstraction. The HMM expresses what the action is like by symbolic representation of time-series data. In this work, we combine information from features that are based on image differences, audio differences, video differences, and motion differences for feature extraction. Hidden Markov models provide a unifying framework for jointly modeling these features. HMMs are used to build scenes from video which has already been segmented into shots and transitions. States of the HMM consist of the various segments of a video. The HMM contains arcs between states showing the allowable progressions of states. The parameters of the HMM are learned using training data in the form of the frame-to-frame distances for a video labeled with shots, transition types, and motion.

3.2 The Second Layer

The semantic concepts process is performed in a hierarchical manner. In hierarchical recognition, a motion data is interpreted at various levels of abstraction. At first, rough recognition is performed and then more detailed recognition is carried out as the process goes down to the lower level. The advantages of using hierarchy are as follows: recognition of various levels of abstraction, simplification of low level models and response to data easily.

At the semantic context level, the proposed fusion schemes that include feature construction and probabilistic modeling take the result from the first level as a basis for characterizing semantic context. The characteristics of each event are then modeled by an HMM in terms of the extracted features from the first level. The results from the event detection are fused by using statistical models: HMM. The fusion work is viewed as a pattern recognition problem, and similar features (detection result of audio events) would be fused to represent a semantic context. See Figure 2.

We use HMMs to characterize different events. For each event, 100 short video clips each 5 – 10s in length are selected as the training data. Based on the results extracted from the training data, a complete specification of HMM with two model parameters (model size and number of mixtures in each state) would be determined.

The HMM-based fusion scheme constructs a general model for each semantic context and tackles different combinations of relevant events. A hidden Markov model $\lambda = (A, B, \pi)$ consists of the following parameters:

1. N, the number of states in the model
2. M, the number of distinct observation symbols in all states
3. the state transition probability distribution
4. the observation probability distribution
5. the initial state distribution

For each semantic context, the parameters of HMM are estimated from the Baum-Welch algorithm by giving sets of training data. The state number N is set at four, and the number of distinct observation symbols M is also four in here.

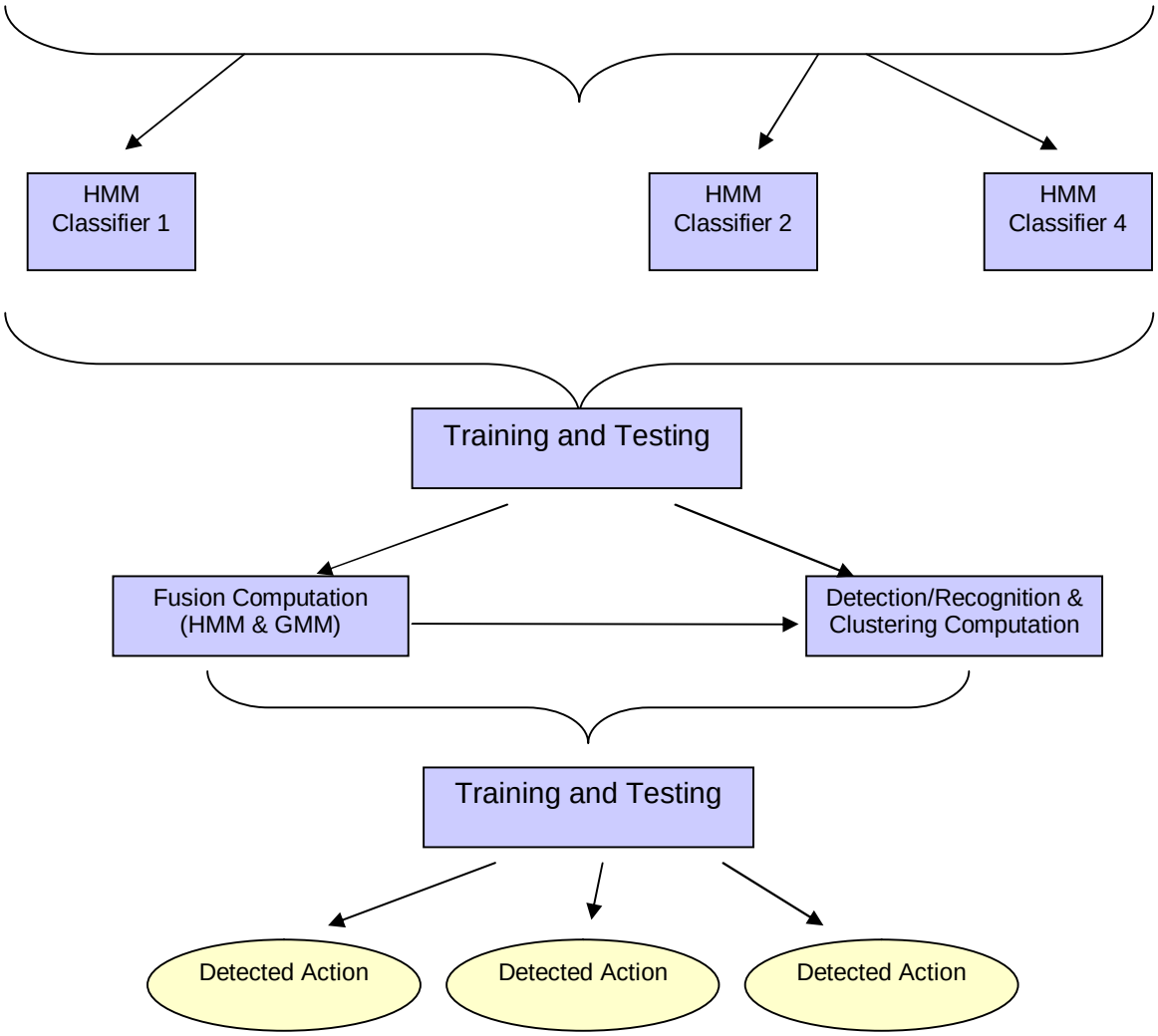

Figure 2: The Schematic Second Layer

Seven HMMs were trained with continuous data extracted from video sequences, one HMM for each action that would be recognized. The HMMs are trained with the data set using the EM algorithm. The parameters of the HMM (including the distribution of each state and transition probabilities between states) are learned using training data of a video manually labelled with shots, transition types, and motion types. Note that in fact, this model is not a “hidden” one, as the states are pre-labelled, and the probability distribution of each state is trained using training segments with the corresponding label. Once the HMM is trained, a given video is segmented into its component shots and transitions, by applying the EM algorithm, to determine the most likely sequence of states. These HMMs are then form the basis for automatically recognize test video sequences into blocks that each relate to one particular action. And since the domain-specific classes jump/run/walk/kick/stand/punch/sit in action movies are very diverse in themselves, a set of video shots for each class were used to capture the structure variations. K=6 with 3-state HMMs topology were trained for jump/run/walk/kick/stand/punch, respectively. The observations are modelled as a mixture of Gaussians. Once HMM models are learned, they can be used to parse new video clips into jump/run/walk/kick/stand/punch/sit action. The data likelihood was evaluated for each of the set of pre-trained model, e.g.

jump models $\theta_j = \{\theta_j^1,...,\theta_j^K\}$ against each feature chunk *F(t)*, to get likelihood values

$$\overline{Q_j^k}(t), k = 1,...,K; t = 1,...,T_\omega. \quad (11)$$

And similarly, for *run* models $\theta_r = \{\theta_r^1,...,\theta_r^K\}$, the likelihood values are

$$\overline{Q_r^k}(t), k = 1,...,K; t = 1,...,T_\omega. \quad (12)$$

The maximum-likelihood values were taken into decision among all HMMs as the label for the current feature chunk. HMM likelihood represents the fitness of each model for every short segment. Although exist a huge body of literature about models of human and biological motion dynamics including data from physiological studies, it is believed that the parameters of the dynamical representations should be estimated from example data. Given the number of linear dynamical systems *M* and the HMM topology, an iterative nonlinear training technique was presented here which enable to estimate the system parameters of each of the action model $\phi_m := [A0_m, A1_m, B_m]$.

4. EXPERIMENTAL RESULTS

The application is implemented using Mathlab Tools with Microsoft Visual C++. All of the tests are performed using Microsoft Windows XP Professionals v.2002 operating system running on HP Workstation xw4100 Pentium ® IV CPU 2.40GHz, 512MB of RAM. In order to determine how well each action is recognized as well as overall detection performance, a measurement method is needed. To evaluate such performance and effectiveness of detection, the following rules and metrics are used:

- Correct and false detection: consider any candidate detection as a correct detection if more than 10% of the actual transition ranges in included in the associated candidate detection range. That is, each correct detection must hit any portion of actual transition with at least 10% temporally overlapped frames.
- Precision and Recall: given a set of correct and false detection, calculate the detection performance by two traditional metrics, precision and recall, which are widely used for performance evaluation in detection and recognition process.

$$\mathrm{Re}call = \left(\frac{\#\ of\ \ correct\ \ \mathrm{det}ections}{\#\ of\ \ actual\ \ action} \right) \quad \text{or}$$

$$\mathrm{Re}call = \frac{CorrectDetection}{TotalShots} \quad (13)$$

$$\mathrm{Pr}ecision = \left(\frac{\#\ of\ \ correct\ \mathrm{det}ections}{\#\ of\ \ \mathrm{det}ections} \right) \quad \text{or}$$

$$\mathrm{Pr}ecision = \frac{CorrectDetection}{CorrectDetection + FalseDetection} \quad (14)$$

Note that these two metrics “globally” indicate how effectively algorithms detect the presence of violence action. An ideal retrieval system is one that retrieves all of the items or information

appropriate to the user's need and which retrieves nothing else; furthermore, it must do so on every occasion. Retrieval effectiveness may be quantified.

This work focuses on using produced data, specifically VCD/DVD movies as a dataset. The rich nature of the VCD/DVD content allows users to use the result of this work in other domains that have domain specific grammar and have specific structural elements (e.g. baseball videos etc.) present.

Movies of different genres were digitized including action, adventure, horror and drama to create a database of a few hours of video. Data from seven movies (Ghost Rider, Matrix, Charlie's Angel, 2 Fast 2 Furious, I Robot, Terminator and Rush Hour) has been used for the experiments. The feature movies cover a variety of genres such as horror, drama, and action. Each input scene contains approximately 20 – 30 shots. To prove the usefulness of the proposed method, few experiments were performed to evaluate the detection performance with several video scenes. For the experiment, ten samples scenes with multi-modal features, *i.e*., *Smultimodal = {S*1, *S*2, *S*3, *S*4, *S*5, *S6, S*7, *S*8, *S9*, *S10}*, were selected to consider dependency among features. The experimental results show that the proposed method to detect simple violence action gives high detection rate and reasonable processing time. The action detection time was calculated for the case with the multimodal feature set. It takes about 102 seconds to detect action within single video clip with PC (Pentium IV CPU 2.40GHz). The overall process time depends on various factors: CPU clock speed, the type of language used in system implementation, optimization scheme, the complexity and the number of processing steps, etc. Because the proposed action detection is performed with unit of short video clip, the detection time is not affected by the length of entire video.

$$Detection\_rate = \sum_{i=r}^{M} \binom{M}{i} P_g^i \left((1-P_g\right)^{(M-i)} \tag{15}$$

where *r = Round(M/2),M = 1, 3, 5, . . . (odd numbers).*

Action detection rate is obtained by equation (15), *M* is the number of video clips chosen from a video and P_g is the probability to classify one action for single video frame. By using processes for motion computation, the result of successful detected action is depicted in Table 1. As is seen, sitting is characterized as the most significant motion with 80% of success rate. This is due to less motion activities involved. These actions were taken from the dataset itself. For example, most of the standing and walking scenes were collected from movie I, Robot. Most of the falling scenes were captured from Rush Hour, sitting and punching from Charlie's Angel, and kicking and running from Matrix.

Type of Sequence	Total Number	Correctly Classified	% Success
Standing	4	3	75
Sitting	5	4	80
Walking	3	2	67
Punching	5	3	60
Falling	4	3	75
Kicking	6	3	50
Running	2	1	50

TABLE 1: Classification of the individual action sequences

The audio test-data set contains six test sets for each event. The database currently contains data on 10 cries, 10 shots, 10 screams, 10 engines and 10 explosions. This experiment series

contained a total of 30 tests. By analyzing and compute the specified audio information (amplitude, frequency, pitch, etc.) needed for classifying audio, Table 2 shows the classification rate. It showed that the classification of audio within a shot was 77.2% in average in the experiment. From table, it shows explosion scored the highest accuracy due to the loudness of it. Screams stands second place as they have the loudness and identified characteristics. Gunshots might have integrated sound so they were not highly recognized. As engines usually combined with fire visual, engine sound is only substitute to this experiment and it always comes in a package of sound (explosion, screeching, bang, etc). These were captured from the movie *2 Fast 2 Furious.*

Audio		**Results in Percent (%)**		
	Correctly classified	**No recognition possible**	**Falsely classified**	**Σ**
Gunshot	81	10	9	100
Cry	51	32	17	100
Explosion	93	7	0	100
Scream	85	10	5	100
Engine	76	9	15	100
Average	**77.2**	**13.6**	**9.2**	**100**

TABLE 2: Audio classification results

Performances were compared over each features, i.e. visual, audio and motion, and performances obtained when features are combined. To compare the usefulness of the proposed multimodal features in this multidimensional framework for action detection, the classification performances of these three cases were evaluated. One case with audio features only, another case with visual features, and the other with audiovisual features. Refer Table 3, Table 4, and Table 5. Table 6 demonstrates a summary of the results for the different methods used in these experiments. Overall, from the table, both the precision rate and recall rate are satisfactory. This experiment gives the best results when audiovisual features are included in the detection process.

SampleVideo	**Correct**	**Miss**	**Fault**	**Precision**	**Recall**
S1	55	15	22	55/77 = 0.714	55/70 = 0.786
S2	14	1	2	14/16 = 0.875	14/15 = 0.933
S3	2	8	3	2/5 = 0.400	2/10 = 0.200
S4	15	5	7	15/22 = 0.682	15/20 = 0.750
S5	11	30	12	11/23 = 0.478	11/41 = 0.268
S6	24	43	19	24/43 = 0.558	24/67 = 0.358
S7	11	17	4	11/15 = 0.733	11/28 = 0.393
S8	10	6	3	10/13 = 0.769	10/16 = 0.625
S9	8	18	3	8/11 = 0.727	8/26 = 0.307
S10	6	16	13	6/19 = 0.316	6/22 = 0.273

TABLE 3: Performance based on audio features

SampleVideo	Correct	Miss	Fault	Precision	Recall
S1	64	6	22	64/86 = 0.744	64/70 = 0.914
S2	10	5	3	10/13 = 0.769	10/15 = 0.667
S3	6	4	2	6/8 = 0.750	6/10 = 0.600
S4	18	2	12	18/30 = 0.600	18/20 = 0.900
S5	28	13	6	28/34 = 0.823	28/41 = 0.683
S6	47	20	9	47/56 = 0.839	47/67 = 0.701
S7	18	10	9	18/27 = 0.678	18/28 = 0.643
S8	13	3	0	13/13 = 1.000	13/16 = 0.813
S9	17	9	4	17/21 = 0.809	17/26 = 0.653
S10	14	8	4	14/18 = 0.777	14/22 = 0.636

TABLE 4: Performance based on colour and motion features (visual feature)

SampleVideo	Correct	Miss	Fault	Precision	Recall
S1	55	15	24	67/79 = 0.848	67/70 = 0.957
S2	15	0	1	15/16 = 0.938	15/15 = 0.100
S3	8	2	1	8/9 = 0.888	8/10 = 0.800
S4	20	0	7	20/27 = 0.741	20/20 = 1.000
S5	34	7	1	34/35 = 0.971	34/41 = 0.829
S6	61	6	2	61/63 = 0.971	61/67 = 0.910
S7	27	1	3	27/30 = 0.900	27/28 = 0.964
S8	13	3	2	13/15 = 0.866	13/16 = 0.812
S9	23	3	1	23/24 = 0.958	23/26 = 0.884
S10	20	2	1	20/21 = 0.952	20/22 = 0.909

TABLE 5: Performance based on audiovisual features

	By audio only (%)		By visual only (%)		By both audio visual (%)	
SampleVideo	Precision	Recall	Precision	Recall	Precision	Recall
S1	71.4	78.6	74.4	91.4	84.8	95.7
S2	87.5	93.3	76.9	66.7	93.8	100
S3	40.0	20.0	75.0	60.0	88.9	80.0
S4	68.2	75.0	60.0	90.0	74.1	100
S5	47.8	26.8	82.3	68.3	97.1	82.9
S6	55.8	35.8	83.9	70.1	97.1	91.1
S7	73.3	39.3	67.8	64.3	90.0	96.4
S8	76.9	62.5	100	81.3	86.7	81.2
S9	72.7	30.7	80.9	65.3	95.8	88.4
S10	31.6	27.3	77.7	63.6	95.2	90.9
Average	**62.52**	**48.93**	**77.89**	**72.10**	**90.35**	**90.65**

TABLE 6: Action detection results by audio only, audiovisual combination

From Table 6, the average recall rate was 90.65% and the average precision rate was 90.35%.**.** As shown in the table, the average recall rate from using audio feature only (48.93%) is lower than those from using visual features only (72.10%) because video data has the complex structure which combines dialogs, songs, colours, and motion activities. The results clearly indicate the effectiveness and the robustness of the proposed algorithm and framework for multidimensional and multimodal styles. Using the audio and visual feature together yielded the highest classification accuracy. As for *S1*, the highest recall rate (95.7%) gained by using both the

audio and visual features. *S1* consists of scenes with fire and action, taken from the movie Ghost Rider.

5. CONCLUSION

In this paper, we present a recognition method of human action that utilizes hierarchical structure. By utilizing hierarchical structure, recognition of various levels of abstraction for one action data, simplification of low level models and response to data by decreasing the level of details become possible. The hierarchical approach will also bridges the gaps between low level features and high level semantics to facilitate semantic indexing and retrieval.

We addressed the problem of modeling and recognizing actions, proposing a two layer HMM framework to decompose action analysis problem into two layers. The first layer maps low level audio visual features into one confidence score. The second layer uses results from the first layer as input to integrate and fusing multiple media clues (audio, visual and motion) to recognize actions.

Now the prototype system is under development. Full experiments results will be put out soon. This work is to believe would be very useful for achieving semantic multimedia retrieval.

6. REFERENCES

1. S. W. Smoliar and H. Zhang, “Content-based Video Indexing and Retrieval”. IEEE Multimedia, pp.62 – 72. 1994.

2. W. Niblack, et al., “Query by Images and Video Content: The QBIC System”. Computer, vol. 28 no. 9, pp. 23 – 32, 1995.

3. S. F. Chang, W. Chen and H.J. Meng, et al., “A Fully Automated Content-based Video earch Engine Supporting Spatio-temporal Queries”, IEEE Trans. Circuits System Video Technology, vol. 2, pp. 602 -615, 1998.

4. J. S. Boreczky and L.D. Wilcox, “A Hidden Markov Model Framework for Video Segmentation using Audio and Image Features”, in Proceedings of the International Conference Acoustics, Speech, Signal Processing, pp. 3741 – 3744, 1998.

5. D M Gavrila. “The Visual Analysis of Human Movement: A Survey”, Computer Vision and Image Understanding, vol. 3 no.1, pp.82 - 98, 1999.

6. S. Fischer, R. Lienhart, and W. Effelsberg, “Automatic Recognition of Film Genres”, Proceedings of ACM Multimedia, pp. 295 – 304, 1995.

7. S. Seitz and C.R. Dyer, “View MorthiMorphing: Uniquely Predicting Scene Appearance from Basis Images”. Proceedings on Image Understanding Workshop, pp. 881 – 887, 1997.

8. J. Yamato, J. Ohya, and K. Ishii, “Recognizing Human Action in Time-Sequential Images using Hidden Markov Models”. Proceedings of Computer Vision and Pattern Recognition, pp. 379 – 385, 1992.

9. K. Sato, J. K. Aggarwal, “Tracking and Recognizing Two-Person Interactions in Outdoor Image Sequences”. Proceedings of IEEE Workshop on Multi Object Tracking, pp. 87 – 94, 2001.

10. S. Hongeng, F. Bremond and R. Nevatia, “Representation and Optimal Recognition of Human Activities”. IEEE Proceedings of Computer Vision and Pattern Recognition, pp. 818 – 825, 2000.