Extracting and Integrating Multimodality Features via Multidimensional Approach for Video Retrieval

Lili N.A.[1], S.A.M. Noah[2] and Fatimah Khalid[1]

[1]Department of Multimedia, Faculty of Computer Science and Information Technology
University Putra Malaysia,UPM Serdang, Selangor
[2]Department of Information Science, Faculty of Technology and Information Science, University Kebangsaan Malaysia, Bangi Selangor.

Summary
This work discusses the application of an Artificial Intelligence technique called data extraction and a process-based ontology in constructing experimental qualitative models for video retrieval and detection. We present a framework architecture that uses multimodality features as the knowledge representation scheme to model the behaviors of a number of human actions in the video scenes. The main focus of this paper placed on the design of two main components (model classifier and inference engine) for a tool abbreviated as VSAD
(Video Action Scene Detector) for retrieving and detecting human actions from video scenes. The discussion starts by presenting the workflow of the retrieving and detection process and the automated model classifier construction logic. We then move on to demonstrate how the constructed classifiers can be used with multimodality features for detecting human actions.
Finally, behavioral explanation manifestation is discussed. The simulator is implemented in bilingual; Math Lab and C++ are at the backend supplying data and theories while Java handles *all front-end GUI and action pattern updating.*

Key words:
Data extraction, multimodality, multidimensional, video retrieval and detection

1. Introduction

Human action is defined as a sequence of body postures with a specific timing constraint. Human actions are characterized by the spatio-temporal structure of their motion pattern. The detection and understanding of human action in videos is of high value for many applications: human computer interaction, surveillance, collaborative environments, training and entertainment, and medical support systems. Human action detection from video is an active research area in recent years. Detecting human actions has been one of the most interesting and challenging problems in multimedia applications. The automatic recognition of human actions has an increased importance in the last years, due to its utility in the surveillance and protection on civil areas.

Manuscript received February 5, 2009
Manuscript revised February 20, 2009

Recognizing human action is a challenging task due to the multiple body parts of interacting persons. First, it involves in getting the whole body motion data. For this, various techniques such as infrared rays can be considered. Then, it involves the interpretation of the human motion, which includes modelling of action, feature extraction, classification, and detection of the action. In general, almost all human action recognition systems work mainly at visual level only, but other information modalities can easily be available, and used as complementary information to retrieve and explain interesting action in a scene.

Our proposed technique characterizes the scenes by integration cues obtained from both the video and audio tracks. 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.

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 [6].

Many approaches have been proposed for behaviour recognition using various methods including Hidden Markov Model, finite state automata, context-free grammar, etc. [7] 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. [8] presented a method to use spatio-temporal velocity of pedestrians to classify their interaction patterns. [9] 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. Create table on comparison recognition percentage with using single source either visual or audio track alone, or combined audio-visual information.

3.1 Determine an appropriate feature to represent the image: Colour Histogram and Histogram Intersection

The distribution of colours in an image has proven to be very useful for object recognition. Building the colour indexes, colour distributions are now an integral part of many recognition schemes. This is not to say that colour alone suffices but rather that colour is one important cue that aids recognition. The colour information, for instance, can be represented using different colour models (e.g., RGB, HSV, YCbCr) and mathematical constructs, such as colour histograms, colour moments, colour sets, colour coherence vectors, or colour correlograms.

Colour histogram counts how much of each colour occurs in the frame. Given a discrete colour space defined by some colour axes (e.g. red, green, blue), the colour histogram is obtained by discretizing the image colours and counting the number of times each discrete colour occurs in the image array. The colours are defined by a normalization process:

$$r = \frac{R}{R+G+B}; g = \frac{G}{R+G+B}; b = \frac{B}{R+G+B} \quad (1)$$

To identify objects based on their colour histograms, we must be able to judge the similarity of the colour histogram of an image to the colour histograms in the database. A method of comparing image and model histograms called Histogram Intersection, which tells how many of the pixels in the model histogram are found in the image. It is especially suited to comparing histograms for recognition because it does not require the accurate separation of the object from its background or occluding objects in the foreground. Experiments show that Histogram Intersection can distinguish models from a large database and accuracy is insensitive to the histogram resolution used.

Given a pair of histograms, I and M, each containing n bins, the intersection of the histograms is defined to be:

$$\sum_{j=1}^{n} \min \left(I_j, M_j\right) \quad (2)$$

The result of the intersection of a model histogram with an image histogram is the number of pixels from the model that have corresponding pixels from the model that have corresponding pixels of the same colour in the image. To obtain a fractional match value between 0 and 1, the intersection is normalized by the number of pixels in the model histogram. The match value is then

$$H(I, M) = \frac{\sum_{j=1}^{n} \min \left(I_j, M_j\right)}{\sum_{j=1}^{n} M_j} \quad (3)$$

3.2 Motion

The motion feature (moments of the motion field, motion histogram, motion intensity, or global motion parameters) is an important characteristic contained in video data. Motion differentiates the dynamic video data from the static imagery data. Most of the past works derived camera motion and operations parameters (zooming, panning, tilting) as their motion signatures. The drawbacks are difficult to support the direct access and manipulation of a specific object of interest. It is best to use object-based motion parameters (rotational and translational) to describe the motion activity.

As motion is an important attribute of video and implicates some semantic cues in visual perception, we employ local motion to compute visual attention value. Motion information can be generated by block matching or optical flow techniques. The HMM expresses what the action is like by symbolic representation of time-series data. Different actions have different motion intensity and show different motion patterns. For instance, chasing cars action has larger motion intensity than reading news.

Spatio-temporal dynamic activity of each video shot from motion sequence:

$$MotionInte\ \ nsity\ (\lambda) = \frac{1}{T}\sum_{i=b+1}^{e}\left(\sum_{m,n}\left|m_i^k(m,n)\right|\right) \quad (4)$$

$m_i^k(m,n)$ = the i^{th} frame of motion sequence within the k^{th} video shot beginning at b^{th} frame and ending at e^{th} frame, and T is the associated shot length ($T = e - b$). The shorter length and more motion each shot has, the higher value its motion density indicates.

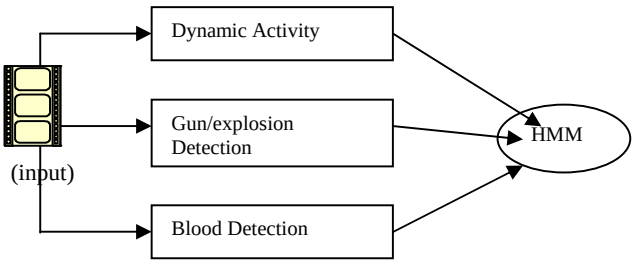

Fig. 1: Block diagram of low visual feature extraction

3.3 Audio

Sound is a powerful tool in filmmaking. As loudness is a fundamental component of film sound, it plays an important role in defining the overall texture of film. We use an audio feature vector consisting of n audio features, which is computed over each short audio clip. There are many audio clip features that are very useful. Getting some of them is time consuming and some of them are not important.

Audio events are defined as short audio clips, which represent the sound of an object or an event/action. Based on elaborately selected audio features, fully connected HMMs will be used to characterize audio events, with Gaussian mixtures modelling for each state. To analyze audio sequences, several audio features from time-domain amplitude and frequency-domain spectrogram are extracted and utilized. In order to achieve computational simplicity and detection effectiveness, we use the below features: volume (energy), zero crossing rate, bandwidth, spectral flux.

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:

$$Vol = \frac{\int_0^{\omega_s} |SF(\omega)|^2 d\omega}{Vol_{max}} \quad (5)$$

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

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.

Spectral Flux / Frequency Centroid (FC) are 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:

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

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:

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

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. The extracted features from each audio frame are concanated as a feature vector. In this experiment, all audio

streams will down sampled to the 16 KHz, 16 bits, and mono-channel format. Each audio will be 25 milliseconds.

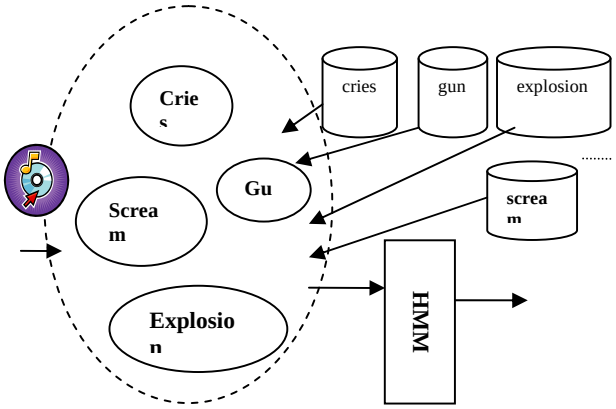

Fig. 2: Block diagram of audio feature extraction

3.4 Fusion

HMM is very effective for capturing the dynamic behaviour of a temporal and time-varying event. It can integrate multimodal features easily. Multimodal interaction can enhance the content findings of one source by using similar content knowledge extracted from the other sources. A discrete HMM is characterized by $\lambda = (A, B, \Pi)$, where A is the state transition probability matrix, B is the observation symbol probability matrix, and Π is the initial state distribution.

We will use multi-stage/multi-dimensional HMM to fuse the multimodal features. Multi-dimensional HMM:

- normal HMM trained on three feature sets: audio, colour and motion;
- multi-stream HMM combining individual audio and visual streams;
- asynchronous HMM combining audiovisual streams.

The lower layer is where a basic HMM is trained on individual modal features. See Fig. 3. Each stream is modelled independently. An audio based HMM classifier (HMM Classifier 1) is first used to separate the input video sequence into few types of sound: scream, cries, explosion, etc. In the second stage, visual based HMM classifiers (HMM Classifier 2) are used to recognize colour of blood, flames and darkness. Same goes to third stage to use motion (HMM Classifier 3) to calculate the intensity of fast or slow motion. Shot break can be detected based on frame differences in both colour histogram and motion field. Separate audio and visual mappings (X) containing different combination rules of different multimodal information. This is where basic HMMs are trained on combined audio-visual features. This method involves aligning and synchronizing audio-visual features to form one concatenated set of features which is then treated as a single stream of data. Then to detect high-level semantic content we will be using pseudo-semantic feature modelling by applying two statistical techniques: GMM and HMM. The final detection is based on the fusion of the outputs of the three modalities by estimating their joint occurrence.

Recognition was done simply by selecting the HMM that was most likely to generate the given sequence of feature vectors. The main advantage of such an approach is that adding a new action can be simply done by training a new HMM.

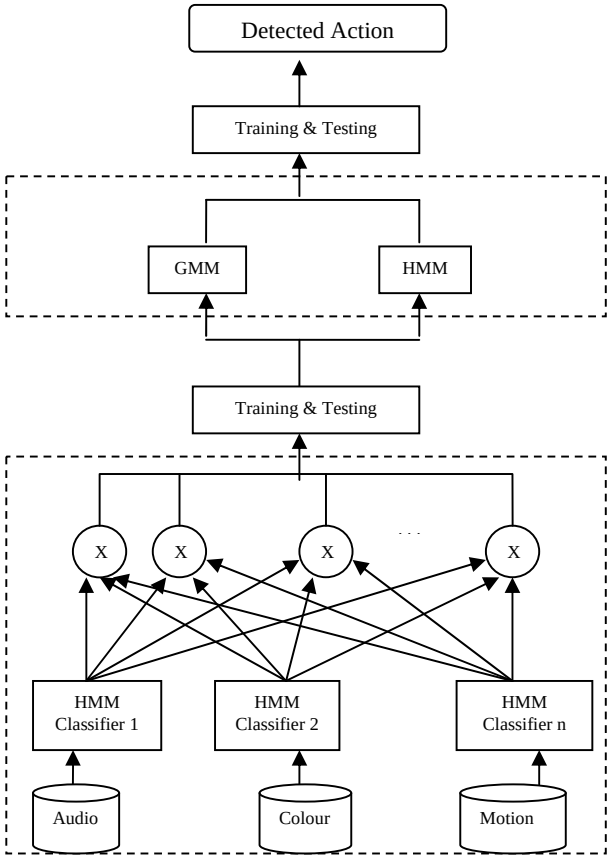

Fig. 3: The Multidimensional System Framework

4. Results and discussion

For each event, short video clips of 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. Each HMM must be trained so that it is most likely to generate the symbol patterns for its category. Training an HMM means optimising the model parameters (A, B, π) to maximise the probability of the observation sequence $P(\theta|\lambda)$. See Fig. 4 to view some samples of collected data.

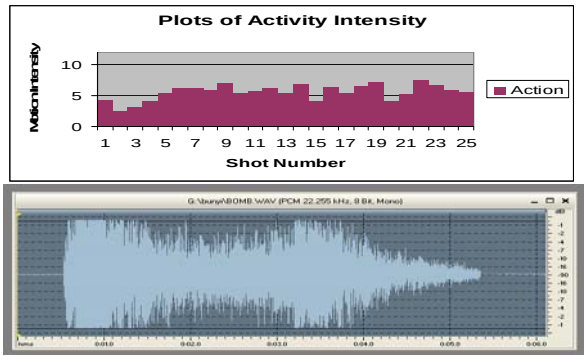

Fig.4: Some samples of collected data on motion and audio

We will be testing our algorithm on action movies. The original frames sequences captured contain complicated background and the positions of the actor/actresses moves during sequence. So, human area extraction and tracking are needed. We will be using the image pre-processing operations described below:

a) preparing background
b) blurring the images with a low pass filter
c) extracting human area
d) binarized the extracted images so that the white and black pixels corresponded to background and human areas

We will be using precision and recall to measure our recognition results, which are well known metrics originally, defined in the information retrieval literature. Precision measures the proportion of correctly recognized actions, while recall measures the proportion of actions that are recognized. The correctness of the detection results and missing detection of the correct actions are judged by humans. We identified the percentage of recall and precision of the proposed framework. See Table 1.

5. Conclusion

We presented an approach to characterize and abstract human activity to support high level video indexing in movie databases. Human activities are represented by combining multiple audio-visual features. We hope that the proposed algorithm detects human actions by detecting spatio-temporal (time space and movement) phenomena which are physically associated with a human action in nature. The use of audio visual information and the adaptive components (HMM) to learn the entire actions represents on the important difference to related works.

Table 1: Action detection results by audio only, audiovisual combination

Feature / *Video Clip*	*By audio only*		*By audio/visual combination*	
	Precision	***Recall***	***Precision***	***Recall***
A	*15/22 = 68.2%*	*15/20 = 75%*	*20/27 = 74.1%*	*20/20 = 100%*
B	*14/16 = 87.5%*	*14/15 = 93.3%*	*15/16 = 93.8%*	*15/15 = 100%*
C	*55/77 = 71.4%*	*55/84 = 65.5%*	*84/89 = 94.4%*	*84/84 = 100%*
D	*51/87 = 58.6%*	*51/73 = 69.9%*	*70/79 = 88.6%*	*70/73 = 95.9%*
E	*11/15 = 73.3%*	*11/28 = 39.3%*	*27/30 = 90.0%*	*27/28 = 96.4%*
F	*75/109 = 68.8%*	*75/98 = 76.5%*	*93/106 = 87.7%*	*93/98 = 94.9%*
G	*40/49 = 81.6%*	*40/93 = 43.0%*	*92/98 = 93.9%*	*92/93 = 98.9%*
AVERAGE	**72.77%**	**66.07%**	**88.93%**	**98.01%**

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 Search 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. Seitz and C.R. Dyer, "View MorthiMorphing: Uniquely Predicting Scene Appearance from Basis Images". *Proc. Image Understanding Workshop*, pp. 881 – 887, 1997.

[7] 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.

[8] 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.

[9] 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.

Lili Nurliyana completed PhD from University Kebangsaan Malaysia in 2008. Currently she is Senior Lecturer at the department of Multimedia, Faculty of Computer Science and Information Technology, University Putra Malaysia. Her research interest includes video retrieval, computer game and animation.

Fatimah Khalid completed PhD from University Kebangsaan Malaysia in 2008. Currently she is Lecturer at the Department of Multimedia, Faculty of Computer Science and Information Technology, University Putra Malaysia. Her research interest include image processing and computer vision.

Shahrul Azman Mohd Noah obtained his bachelor of science degree in Mathematics at the Univesiti Kebangsaan Malaysia in 1992. Having spent a year working in Cellular Communications Network (CELCOM) as a research and planning executive he then pursued his MSc and PhD studies at the Department of Information Studies, University of Sheffield (1994 - 1998). He is currently an associate professor at the Department of Information Science, Faculty of Information Science & Technology, Universiti Kebangsaan Malaysia. His research interests include semantic web, conceptual modeling, data warehouse design, knowledge representation and agent based information retrieval.