A Framework for Human Action Detection via Extraction of Multimodal Features

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

Abstract

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 VASD (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. 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, human action detection, multimodal, hidden markov model.

1. INTRODUCTION

Action is the key content of all other contents in the video. Action recognition is a new technology with many potential applications. Action recognition can be described as the analysis and recognition of human motion patterns. 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. 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 that analyze audio and video data separately, this research presents a method which able to integrate audio and visual information for action scene analysis. The approach is top-down for determining and extract action scenes in video by analyzing both audio and visual data. A multidimensional layer framework was proposed to detect action scene automatically. 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. With this framework, the gaps between low-level features and the semantic contexts that last in a time series was bridged. The experimental evaluations indicate that the approach is effective in detecting high-level semantics such as action scene.

2. RELATED WORKS

The problem of human action recognition is complicated by the complexity and variability of shape and movement of the human body, which can be modelled as an articulated rigid body. Moreover, two actions can occur simultaneously, e.g., walk and wave. Most work on action recognition involving the full human body is concerned with actions completely described by motion of the human body, i.e., without considering interactions with objects.

Previous works on audio visual content analysis were quite limited and still at a preliminary stage. Most of the approaches are focused on visual information such as colour histogram differences, motion vectors and key frames [3,4,5]. Colour histogram [7] difference and motion vector between video frames or objects are the most common features in the scene recognition algorithms. Although such features are quite successful in the video shot segmentation, scene detection based on such visual features alone poses many problems. The extraction of relevant visual features from an images sequence, and interpretation of this information for the purpose of recognition and learning is a challenging task. An adequate choice of visual features is very important for the success of action recognition systems [6].

Because of the different sets of genre classes and the different collections of video clips these previous works chose, it is difficult to compare the performance of the different features and approaches they used.

Automatic interpretation of human action in videos has been actively done for the past years. The investigation of human motion has drawn a great attention from researches in computer vision or computer graphics recently. Fruitful results can be found in many applications such as visual surveillance. Features at different levels have been proposed for human activity analysis. Basic image features based on motion histogram of objects are simple and reliable to compute (Efros et al. 2003). In (Stauffer & Grimson 2000), a stable, real-time outdoor tracker is proposed, and high-level classifications are based on blobs and trajectories output from this tracking system. In (Zelnik-Manor & Irani 2001), dynamic actions are regarded as long-term temporal objects, and spatio-temporal features at multiple temporal. There is a huge body of literature on the topics of visual tracking, motion computation, and action detection. As tools and systems for producing and disseminating action data improve significantly, the amount of human action detection system grows rapidly. Therefore, an efficient approach to search and retrieve human action data is needed.

3. METHODOLOGY

The process of video action detection is depicted in Figure 1.0.

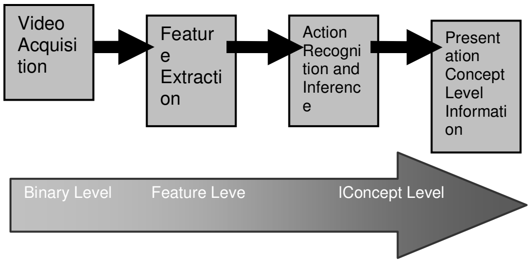

Figure 1.0 Metadata Generation Process

There are three classes of information. The main type of information, which serves as the input of the complete system, is the *binary level information, which* comprises the raw video files. This binary level information is analyzed which results in the so called *feature level information*, i.e. features interesting for detecting certain action like the measurements of action speeds. Finally, from this feature level information, actions are detected that are regarded as the *concept level information*. When a user searches information from the multimedia, it does not have to browse the binary level video files anymore, but it can directly query on concept level. The user can for example query an action in which a man is punching another man, which might cause a violent later on. As a result a more advanced search engine is created that can be used for human action purposes.

In this paper, the feature extraction process (Figure 2.0) is discussed: converting the binary level information coming from the video acquisition system into feature level information suited for further processing. From the flowchart (Figure 3.0), we have four levels of processes:

- Pixel level represents the average percent of the changed pixels between frames within a shot
- Histogram indicates the mean value of the histogram difference between frames within a shot
- Segmentation level to indicate background and foreground areas
- Object tracking based on observation: flames, explosion, gun

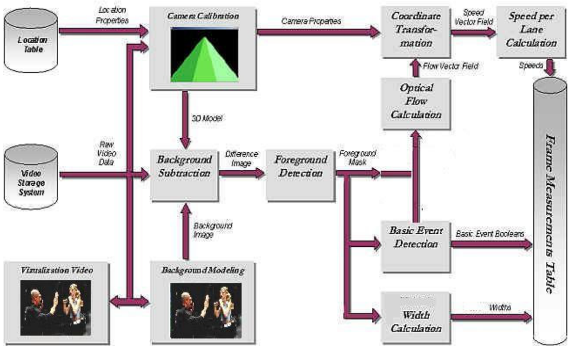

Figure 2.0 The overview of the system

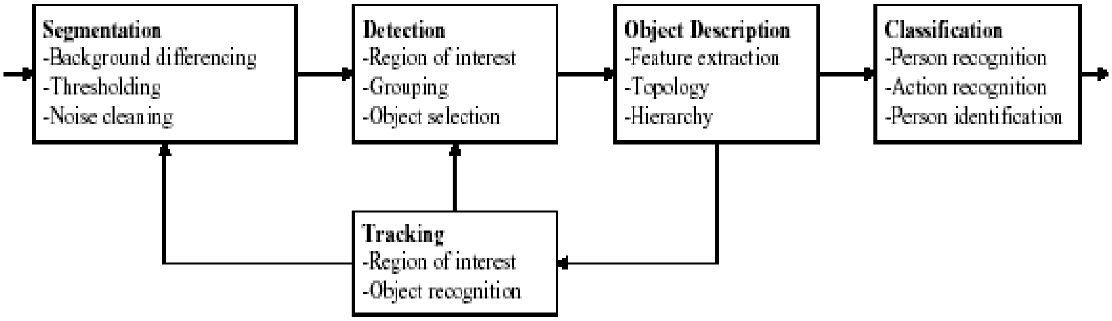

Figure 3.0 Flow chart of feature extraction process

We proposed an action classification method with the following characteristics:

- Feature extraction is done automatically;
- The method deals with both visual and auditory information, and captures both spatial and temporal characteristics;
- Edge feature extraction
- Motion
- Shot activity that conveys large amount of information about the type of video
- Colour feature extraction
- Sound
- The extracted features are natural, in the sense that they are closely related to the human perceptual processing.

3.1 Segmentation

The segmentation and detection stages are often combined and simply called object detection. Basically the task of the segmentation is to split the image into several regions based on color, motion or texture information, whereas the detection stage has to choose relevant regions and assign objects for further processing.

We follow the assumption that the video sequence is acquired using a stationary camera and there is only very little background clutter. Based on this we use background differencing followed by threshold to obtain a binary mask of the foreground region. In order to remove noise median filtering and morphological operations are used. Regions of interest (ROI) are detected using boundary extraction and a simple criterion based on the length of the boundary. Boundary filling is applied to each ROI and the resulting binary object masks are given to the description stage.

3.2 Object Descriptor

The object description refers to a set of features that describe the detected object in terms of color, shape, texture, motion, size etc. The goal of the feature extraction process is to reduce the existing information in the image into a manageable amount of relevant properties. This leads to a lower complexity and a more robust description. Additionally, the spatial arrangement of the objects within the video frame and as related to each other is characterized by a topology.

A very important issue for the performance of a subsequent classification task is to select a suitable descriptor that expresses both the similarity within a class and the distinctions between different classes. Since the classifier strongly depends on the information provided by the descriptor it is necessary to know its properties and limitations relating to this specific task. In case of human body posture recognition it is obvious that shape descriptors are needed to extract useful information.

3.3 Classification

Classification is a pattern recognition (PR) problem of assigning an object to a class. Thus the output of the PR system is an integer label. The task of the classifier is to partition the feature space into class-labeled decision regions. Basically, classifiers can be divided into parametric and non-parametric systems depending on whether they use statistical knowledge of the observation and the corresponding class. A typical parametric system is the combination of Hidden Markov Model (HMM) and Gaussian Mixture Model (GMM), which assumes Gaussian distribution of each feature in the feature vector.

3.4 Posture and Gesture

Based on our overall system approach we treat the human body posture recognition as a basic classification task. Given a novel binary object mask to be classified, and a database of samples labeled with possible body postures, the previously described shape descriptors are extracted and the image is classified with a chosen classifier. This can be interpreted as a database query: given a query image, extract suitable descriptors and retrieve the best matching human body posture label. Other similar queries are possible, resulting in a number of useful applications, such as skeleton transfer, body posture synthesis and figure correction.

3.5 HMM and GMM

An approach for action recognition by using Hidden Markov Model (HMM) and Gaussian Mixture Modelling (GMM) to model the video and audio streams respectively will be proposed. HMM will be used to merge audio-visual information and to represent the hierarchical structure of the particular violence activity. The visual features are used to characterize the type of shot view. The audio features describe the audio events within a shot (scream, blast, gun shots). The edge, motion (orientation and trajectory) information is then input to a HMM for recognition of the action.

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.

For each type of violent activity (punch and kick), a HMM will be build to characterize the action and interaction processes as observation vectors. A HMM for each action is trained with the corresponding training MPEG video sequences (kick, punch, run, walk, stand, etc.). The expectation maximization (EM) and GMM approaches are then used to classify testing data using the trained models. Once the mean and covariance of the Gaussian model of the training audio data are obtained, the likelihood ratio between the input audio track and the sound classes, is computed to determine which class the associated sound belong to.

4. Experimental Discussion

By using these processes for action detection, 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.

TABLE 1 Classification of the individual action sequences

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

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. Figure 4 shows some scenes that demonstrate these actions for classification.

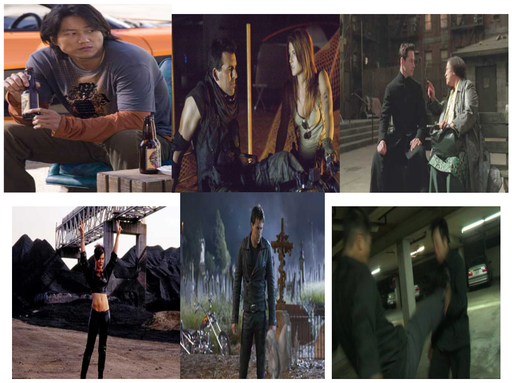

Figure 4.0 Some examples of video clips

5. Conclusion

In this research, we studied on techniques for extracting meaningful features that can be used to extract higher level information from video shots. We will use motion, colour, and edge features, sound and shot activity information to characterize the video data.

The proposed algorithm is implemented in C++ and it works on an Intel Pentium 2.56GHz processor. As described above HMMs are trained from falling, walking, and walking and talking video clips. A total of 64 video clips having 15,823 image frames are used. Some image frames from the video clips are shown in Fig. 4. In all of the clips, only one moving object exists in the scene**.**

In summary, the main contribution of this work is the use of both audio and video tracks to decide an action in video. The audio information is essential to distinguish an action from a person rather than a person simply sitting down or sitting on a floor. To prove the usefulness of the proposed method, the experiments were performed to evaluate the detection performance with several video genres. The experimental results show that the proposed method to detect action scenes 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.

6. References

1. L. Zelnik-Manor and M. Irani, “Event-based Analysis of Video”. *Proceedings of IEEE Conference Computer Vision and Pattern Recognition*, 2001.
2. C. Stauffer and W.E.L. Grimson, “Learning Patterns of Activities using Real-Time Tracking”. *Journal of IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol (22), no. (8), pp. 747 – 757, 2001.
3. A.A. Efros, A.C. Berg, G. Mori and J. Malik, “Recognizing Action at a Distance”. *Proceedings of International Conference on Computer Vision*, 2003.
4. S. Fischer, R. Lienhart and W. Effelsberg, “Automatic Recognition of Film Genres”, *Proceedings of ACM Multimedia*, pp. 295 – 304, 2003.
5. G. Tzanetakis and P. Cook, “Musical Genre Classification of Audio Signals”, *IEEE Trans. On Speech and Audio Processing*, vol. 10, no. 5, pp. 293 – 302, 2002.
6. C.P., Tan, K.S. Lim, and W.K. Lai. 2008. Multi-Dimensional Features Reduction of Consistency Subset Evaluator on Unsupervised Expectation Maximization Classifier for Imaging Surveillance Application. *International Journal of Image Processing*, vol. 2(1), pp. 18-26.
7. J. P and P.S. Hiremath. 2008. Content Based Image Retrieval using Color Boosted Salient Points and Shape features of an image. 2008. *International Journal of Image Processing*, vol. 2(1), pp. 10-17.