A Video Content Classification Algorithm Applying to Human Action Recognition

Tanfeng Sun[1,2], Xinghao Jiang[1,2], Chengming Jiang[1], Yaqing Li[1]
[1]*School of Electronic Information and Electrical Engineering, Shanghai Jiao Tong University, Shanghai 200240, P.R. China*
[2]*National Engineering Lab on Information Content Analysis Techniques, GT036001, Shanghai 200240, P.R. China, phone: 86-21-34206657*
xhjiang@sjtu.edu.cn

***Abstract*—A new classification algorithm of human action recognition of video content is suggested in our paper. It analyzes the variation of the content of video scenes or human action from video bi-modal features, in order to cognize content efficiently and precisely. This scheme is based on the pattern analysis of spatio-temporal slices and audio signature feature extracted from the video files. The Spatial-temporal Variation Histogram feature is firstly defined in our paper. It is applied to describe spatio-temporal variation analysis of human action or video scenes. The audio signature is also applied to identify the audio content by extracting unique signatures from a part of audio signal. The experiments show excellent performance of classification on the KTH dataset.**

***Index Terms*—Image recognition, pattern analysis, audio signature, Spatial-temporal Variation Histogram.**

I. INTRODUCTION

Today a tremendous amount of videos are accessible on the Internet. Due to the large video quantity, it is not easy for a viewer to find out the class of video of interest directly. Requirements to narrow viewers' choice grow and ask for related technology to categorize videos automatically. Although in the past decades, researches for image processing have been flourishing with the developments of computer science and the Internet, theories for manipulating video data have been generated and applied later. Some researchers extended and applied methods for text processing and image processing to video processing through dealing with text feature or key-frames from the video shots.

Generally, a large number of approaches have been attempted to perform automatic semantic cognition of video, which could be divided into text-based approaches, audio-based approaches and visual-based approaches. Text-only approaches are the least common method, which extract viewable text such as text on objects or text placed on screen by optical character recognition (OCR) as text features. Transcript of the dialog is another text feature extracted from speech using speech recognition methods or is provided in the form of closed captions or subtitles. Audio-only approaches are found more often than text-only approaches. They can be divided into time domain features and frequency domain features. Root mean square (RMS) and zero crossing rates (ZCR) are examples of the former while the energy distribution and bandwidth are among the latter. Visual-based approaches are the most accepted method because human receive most information from sense of vision from video. Global feature about color, texture, or high level descriptors such as Scale Invariant Feature Transform (SIFT) are generated from key-frames. However, these methods neglect the important information which videos contain.

The biggest differences between images and videos exist in two aspects. One of them is the information of audio, since the auditory information reflects some unique content of video differing from the image sequences. Another distinction is the relationship of the image sequences, because the static images such as the key-frames of a shot do not have the information of the fore-and-aft frames. Therefore, analyses both of audio and the spatio-temporal features have been proposed since then. For the aspects of audio, both time and frequency domain of the signal are focused to extract descriptors, and most of them have been defined in Moving Pictures Experts Group (MPEG) as a standard.

Simultaneously, the spatio-temporal features developing in recent years are used to recognize human behaviours and actions. Real spatio-temporal processing should exploit the fact that many interesting events in a video sequence are characterized by strong variations of the data in both the spatial and temporal dimensions. Recently, spatio-temporal feature has drawn attention for human action recognition and content-based video analysis [1], [2]. The key idea of spatio-temporal feature is to extend two dimensional features to the temporal dimension. Dollar et al. [3] proposed a method to apply 2-D Gaussian kernels in the spatial space and 1-D Gabor filters in the temporal direction to detect local cuboids. Laptev et al. [4] proposed an extended Harris detector to extract spatio-temporal features while Klaser et al. [5] proposed three-demensional HoG to represent spatio-temporal features. However, these methods are all based on "cuboids" where local cubic spatio-temporal regions are

Manuscript received March 19, 2012; accepted May 29, 2012.
This work was supported by the National Natural Science Foundation of China (No. 61071153), Sponsored by Program for New Century Excellent Talents in University (NCET-10-0569).

extracted. As a result, computational costs to extract cuboids features by the methods described above are relatively high. To overcome these problems, in this paper, we propose a grid of key lines to generate the spatio-temporal slices to create a fast detector. Since we do not use cuboids, the proposed method is more simple and efficient in order to fit for the large amount of video data.

In this paper, a novel scheme is proposed to associate both the greatly special features of video other than static image to evaluate the variation of video scenes. Descriptors which can be effectively express the variation in audio and video are extracted in our method. The influence of these variations to classify video scenes into certain genres is evaluated in our experiments. We apply our method to KTH action recognition dataset.

II. A BI-MODAL SEMANTIC COGNITION SCHEME DESCRIPTION

A. Framework overview

The framework of our method is shown in Fig. 1. It mainly consists of 3 portions.

The first one is the pre-processing of the input videos, which concentrate on the standardization and shot boundary generation because both the audio and spatio-temporal descriptors are extracted from a certain period of shot.

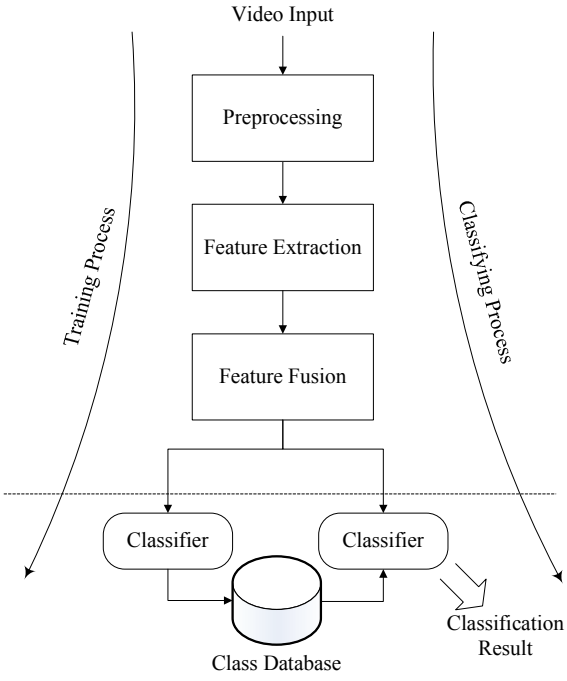

Fig. 1. The Framework of Bi-model semantic cognition model.

The second part of our method refers to feature extraction, thus, in that section, the descriptors are generated. To evaluate audio variation, we firstly separate the audio track and obtain a feature named Audio Signature according to MPEG-7. While to evaluate the visual one, spatio-temporal slices of a shot are firstly made by using a nonlinear cutting method which will be discussed later and then the structure tensor is formed.

Later we define the method to get the tensor histogram as a descriptor. The feature processing, as the last part of our method, includes training and classifying using a SVM classifier, whose training and testing data are with regard to some public datasets.

B. Audio feature extraction

High level descriptors which sever as audio fingerprint, enabling automatic identification of audio content by extracting unique signatures from a period audio signal, are also the part of MPEG-7 audio descriptors. The Audio Signature (AS) is used in our method as the main descriptor to represent the feature of each shot of audio.

To form AS, Audio Spectrum Flatness (ASF) should be calculated at first. Let P(k, bl) represent the spectrum power of the frame k of the input audio signal, ASF is the ratio between the geometric mean and the arithmetic mean of the spectral power within this band

$$ASF(b) = \frac{\sqrt[b_{\max} - b_{\min}+1]{\prod_{bl=b_{\min}}^{b_{\max}} P(k, bl)}}{\frac{1}{b_{\max} - b_{\min} + 1} \sum_{bl=b_{\min}}^{b_{\max}} P(k, bl)}, \quad (1)$$

where b_{max} and b_{min} is the maximum and minimum band of the current band bl respectively.

C. STVH feature extraction

The STVH feature is defined firstly in our scheme. In each shot of a video, we sequentially arrange the frames (X, Y) one by one, so as to construct a 3-D volume with the spatial image according to the time (T). Then, we reconstruct 2-D images by using the pixels in dimension T and another dimension X or Y as a spatio-temporal slice. A Gaussian filter is used to convolute the slice image with an 11 * 11 discrete Gaussian kernel. Image of spatio-temporal slice reflects object motions as shown in Fig. 2.

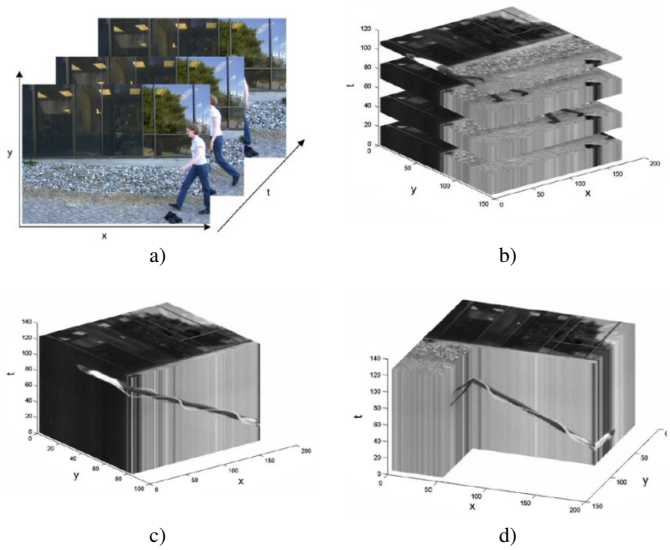

Fig. 2. Representation of a video sequence as a spatio-temporal volume and three different views; a) the original video sequences; b) specific sample frames are combined and reconstructed into a space cube; c), d) different views of spatio-temporal views.

It offers abundant visual cues of a pixel line in the video, which can be exploited for motion content estimation as well

as variation extent evaluation.

While we have assign the key lines of the window, for each key line, we get the spatio-temporal slices, and for every non-margin points in each slices, we calculate the direction of the local structure as well as the variation value and sum the variation values into the direction bins as the descriptor vector V = (v1, v2, …, v12)

$$V_i = \sum_{n,x,t} C(n,x,t), \quad if\ \varphi(n,x,t) \to \varphi_i . \quad (2)$$

A confidence factor is added into the vector which is expressed as

$$V_{13} = \frac{1}{NumberOfFrames} \sum_{i=1}^{12} V_i . \quad (3)$$

To unify the descriptor vector as a normalized one, we find the maximum value of the vector and divide each value by that maximum value.

When an input video comes, we deal with it as several independent segments according to the shot, and then cut the whole video based on the grid of key lines. In this paper, from every shot of video, we get 18 spatio-temporal slices. In each slice, the STVH is calculated according to the algorithm by formula (3). Finally the STVH results of each slice are accumulated as a whole descriptor to represent this video shot.

D. Semantic cognition classifier

SVM learning is used to train representative models for each action and general semantics categories, using STVH and AS as feature vectors. Separate SVM classifiers are trained for each action. The testing video is classified by the classifier which has the greatest distance from the SVM hyper-plane. When the number of the categories is certain one and each semantics are highly similar, we train each semantic separately and applied the DAGSVM method to restrain each video into only one category.

III. EXPERIMENTS

In our experiments, three kinds of videos build up the collection of the test data set. KTH human action dataset from Royal Institute of Technology in Sweden containing 600 videos is select as one of the test data set. These videos, with 6 types of human actions referring to walking, jogging, running, boxing, hand waving and hand clapping, are over homogeneous backgrounds with a static camera and have a length of 20 seconds in average without audio track.

FFMPEG library is applied in our experiment to extract audio stream from the video. Another tool we used is MPEG-7 audio encoder, which can extract MPEG-7 audio features to build our audio effect descriptor. To the classifier, we choose "libsvm" which is commonly used in data cogniton. Since we only focus on the efficiency and effectiveness of the descriptors presented in this paper, the capability of classifier and model is out of discussion. We should notice that we test all the data in the same environment including the classifier.

A. Time cost

We test the computational time cost of the extraction of the variation histogram descriptor at first. This testing is among several two dimensional and three dimensional descriptors. Since the algorithm of variation histogram combines the factors both in time dimension and space dimension, it is necessary to figure out if it works efficiently.

TABLE I. TIME COST COMPARISON.

Descriptor	Time Cost (s)
3-D SIFT	101
Space-time SURF	6.25
2-D SIFT	4.65
2-D SURF	1.55
STVH	0.97

The time costs are shown in the Table I. The time cost of STVH is less than four others methods. We choose a video containing about 500 interest points for descriptors like 3-D SIFT and so on. However, in our method, the time cost is completely independent on the interest points. In another word, whether the pictures of the test video are complicated or not, the time cost of a certain period of the frame sequences remains a constant. This kind of advantage can be applied to real time application due to its time-invariance of the input content.

B. The Comparison of three schemes

We test the influence of STVH on the semantic cognition of videos content on the KTH human action dataset. The sample videos contain 6 types of human actions performed several times by 25 subjects in 4 different scenarios as illustrated in Fig. 2 action types include walking, jogging, running, boxing, hand waving and hand clapping while 4 scenarios are S1 (outdoors), S2 (outdoors with scale variation), S3 (outdoors with different clothes) and S4 (indoors). Thus, the count of videos is: 25(subjects)*6(types of action)*4(scenarios) = 600.

We get a confusion matrix of our semantic cognition as the following shows. The diagonal rank is correct cognition percentage and the others are wrong cognition percentage in confusion matrix.

Table II to Table IV shows the confusion matrix for these six actions of each method. The trace of the matrix is a measure to the semantic cognition of human action, ranging from 100 (expected value for random cognition) to 600 (for perfect cognition). The diagonal rank sum of our matrix is 471.6 in Table IV, higher than 377.8 and 430.3 in Table II and Table III.

TABLE II. CONFUSION MATRIX USING VOLUMETRIC FEATURES [6].

	Box	Clap	Wave	Jog	Run	Walk
Box	69.4	11.1	5.6	2.8	11.1	0.0
Clap	36.1	55.6	2.8	0.0	5.6	0.0
Wave	2.8	0.0	91.7	5.6	0.0	0.0
Jog	0.0	0.0	0.0	36.1	33.3	30.6
Run	0.0	27.8	0.0	25.0	44.4	2.8
Walk	0.0	8.3	0.0	11.1	0.0	80.6

TABLE III. SCHULDT'S LF + SVM CONFUSION MATRIX [7].

	Box	Clap	Wave	Jog	Run	Walk
Box	97.9	0.7	0.7	0.0	0.0	0.7
Clap	35.4	59.7	3.5	0.0	0.0	1.4

	Box	Clap	Wave	Jog	Run	Walk
Wave	20.8	4.9	73.6	0.0	0.0	0.7
Jog	0.0	0.0	0.0	60.4	16.7	22.9
Run	6.3	0.0	0.0	38.9	54.9	22.9
Walk	0.0	0.0	0.0	16.2	0.0	83.8

TABLE IV. CONFUSION MATRIX OF SVH TEST.

	Box	Clap	Wave	Jog	Run	Walk
Box	91.7	5.6	2.8	0.0	0.0	0.0
Clap	20.0	73.0	7.0	0.0	0.0	0.0
Wave	1.0	2.0	97.0	0.0	0.0	0.0
Jog	0.0	0.0	0.0	54.5	26.4	19.1
Run	0.0	0.0	0.0	18.2	73.6	8.2
Walk	0.0	0.0	0.0	14.5	3.6	81.8

We compare our scheme with two other schemes on the KTH human action dataset in Table V.

TABLE V. THREE SCHEME'S COMPARISON EXPERIMENTS ON THE KTH.

Performance	Average Accuracy
Scheme[6]	63.0%
Scheme[7]	71.1%
Our Scheme	78.6%

The average accuracy of our scheme is 78.6%, higher than 63.0% and 71.7% in Table II.

From above discussed, we can conclude that the STVH feature is more efficiently to cognize motion saliency semantics than motionlessness ones.

IV. CONCLUSIONS

In this paper, we have proposed an effective scheme to Human Action of video content with STVH feature and Audio signature feature. A Spatio-temporal Variation Histogram descriptor is defined which is extracted from spatio-temporal slices. The Audio Signature is adopted to combine with STVH descriptor. The combined feature is applied to a novel cognition model framework in our paper. Experimental results the STVH feature is useful to cognize human action in the KTH dataset.

REFERENCES

[1] C. A. Mokhber, M. Milgram, "Recognition of Human Behavior by Space-time Silhouette Characterization", *Pattern Recognition Letters,* vol. 29, no. 1, pp. 81–89, 2008. [Online]. Available: http://dx.doi.org/10.1016/j.patrec.2007.08.016

[2] B. L. Chen, W. H. Chen, S. H. Lin, W. Y. Chu, "Robust Speech Recognition using Spatial-Temporal Feature Distribution Characteristics", *Pattern Recognition Letters,* vol. 32, no. 1, pp. 919–926, 2011. [Online]. Available: http://dx.doi.org/10.1016/j.patrec.2011.01.016

[3] P. Dollar, G. Cottrell, S. Belongie, "Behavior Recognition via Sparse Spatiotemporal Features", in *Proc. of Surveillance and Performance Evaluation of Tracking and Surveillance,* 2005, pp. 65–72.

[4] I. Laptev, T. Lindeberg, "Local descriptors for Spatio-Temporal Rcognition", in *Proc. of IEEE International Conference on Computer Vision,* 2003, pp. 91–103.

[5] A. Klaser, M. MarszaÃlek, C. Schmid, "A Spatio-Temporal Descriptor based on 3Dgradients", in *Proc. of BMVA British Machine Vision Conference,* 2008.

[6] K. Yan, S. Rahul, H. Martial, "Efficient Visual Event Detection using Volumetric Features", in *Proc. of the IEEE International Conference on Computer Vision,* vol. 1, 2005, pp. 166–173.

[7] C. Schuldt, I. Laptev, B. Caputo, "Recognizing human actions: A local SVM approach", in *Proc. of the ICPR,* 2004, pp. 32–36.