Music video classification using audio and visual features

Abstract:- *The main aim of this study is therefore to construct music video classification model that identifies music type and dance style from video clip based on audio and visual features. To build this model we use 100 video clips and then the audio part is ripped from the video. The video part is also segmented to smaller video clips of containing important part of the dance with frame rate of 10 fps. This sequence of image frames is given as an input to log Gabor filter to detect edges and extract features of dancers, since edges appear in the frequency domain as high frequencies. The result shows that the use of both visual and audio feature provides on the average 86.20% accuracy.*

Keyword: *Music and Dance, Content-Based Video Classification, Combined Audio and Visual Features.*

Today the internet is flooded with video of all types like movies, songs, personal videos, etc [1,3]. This is because of the rising of internet bandwidth speeds, the rapid development of digital communication and remote sensing technologies, as well as an increase in computer capacities. In Ethiopia also the number of videos, especially music video clip produced and uploaded over the internet is growing from day-to-day [4,8]. Music has an important social and cultural impact on the people of country for a long period of time. For instance, Ethiopian music and dance style is different from other countries of the world.

Basically, the music video clip captures both audio and visual modalities. The visual contains the information about the appearance (color, texture and shape) [17] and the audio may contain music or speech. Video classification based on audio and visual features is a research area listed under the problem related to computer vision [10,12], and computer audition [15].

The process of video classification consists of many steps, such as pre-processing, segmentation, feature extraction and classification for both audio and visual modalities [3,8,11]. In this study an attempt is made to adopt video classification techniques for categorizing Ethiopian Nations music and dance from a music video source.

To build this model we use 100 video clips and then the audio part is ripped from the video using free open source audio extractor tool. After the audio part is extracted, since all regions in the audio is not necessary for the study we segment the audio clip into smaller 1.5 second audio length frames and then 18 audio features are extracted from this audio frames.

The video part is also segmented to smaller video clips of containing important part of the dance using a free Virtual Dub tool with frame rate of 10 fps. From each video clip we get 60 sequence of image frames. This sequence of image frames is given as an input to log Gabor filter to detect edges and extract features of dancers, since edges appear in the frequency domain as high frequencies, and it is natural to use Log-Gabor filter to pick edges. PCA is then applied for dimensionality reduction and select principal features [11].

The model used in the video classification is shown below.

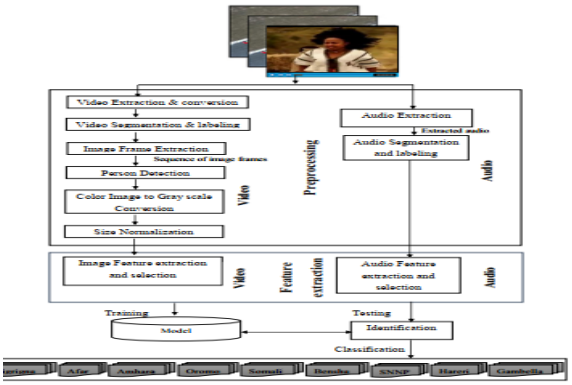

Figure:1.1 Architecture of the proposed content-based classification for Music Video Clip.

Training and testing has been conducted with three experiments; based on the type of features in video clips. The first experiment is conducted using audio frames or audio features while the second experiment is based on the sequence of image frames or visual features. The later experiment is conducted with two phases, with 30 and 60 sequences of image

frames. The third experiment is conducted based on the combination of sequence of image frames and audio frames. The classification result obtained from these three experiments are presented below in a table.

List of experiments	Features used in the experimentation	Classification performance
Experiment 1	Audio features	
	Phase 1: With 1.5 second audio length frames	81.17%
Experiment 2	Visual features	
	Phase 1: With 30 sequence of image frames	82.18%
	Phase 2: With 60 sequence of image frames	82.59%
Experiment 3	Combining audio and Visual features	86.20%

Table 1.1 Summary of experimental results

A multi-class classification SVM, called LIBSVM is used for constructing the classification model. The training and testing set is separated using 5-fold cross validation. The result shows that the use of both visual and audio feature provides on the average 86.20% accuracy.

The use of rhythm for audio feature representation and body shape for visual feature makes our study very difficult. In addition people detection in complex backgrounds and shortage of video resources are also another challenge in our study. Our recommendation in this study is to use better techniques for video segmentation, person detection and also more audio and visual feature extraction for better video representation.

Reference

[1]. D. Brezeale and D. J. Cook, "Automatic video classification: A survey of the literature", IEEE Trans. Syst., vol. 38, no. 3, pp. 416–430, 2008.

[2]. S. Prashant, S. Nirmit, Y. Alok, S. Aseem and K. Ankit, “A Survey on Classification of Videos using Data Mining Techniques", IJCA Special Issue on Issues and Challenges in Networking, Intelligence and Computing Technologies, vol. 5, no. 5, pp. 27–32, 2012.

[3]. M. Roach, J. Mason, E. Nicholas , L. Xu and F. Stentiford, “Recent trends in video analysis: a taxonomy of video classification problems,” B Texact Technologies - Research, vol. 1, no. 1, pp. 348--353, 2002.

[4]. T. Teffera, "The Role of Traditional Music Among East African Societies: The Case Of Selected Aerophones", 16 th International meeting on Folk Musical Instruments , Martin Luther University Halle-Wittenberg, Germany, pp. 1, 2006

[5]. Z. Liu, Y. Wang, and T. Chen, “Audio feature extraction and analysis for scene segmentation and classification,” Journal of VLSI Signal Processing Systems, vol. 20, no. 1-2, pp. 61–79, 1998.

[6]. M. Roach and J. Mason, “Classification of video genre using audio,” International Journal of speech, vol. 4, pp. 2693–2696, 2001.

[7]. S. Jadhav, M. Joshi and J. Pawar “Towards Automation and Classication of BharataNatyam Dance Sequences”, Indian Journal of Science, Vol. 1 pp 321-328, 2010

[8]. T. Fitsum, “Ethiopian Traditional Dance Video Classification Using Visual Features,” Msc Thesis, University of Gondar, Gondar, Ethiopia, 2014

[9]. J. Huang, Z. Liu, Y. Wang, Y. Chen, and E. K. Wong, “Integration of multimodal features for video scene classification based on HMM,” IEEE Workshop on Multimedia Signal Processing, Vol 1, pp. 53-58, 1999

[10]. N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection”, in Computer Vision and Pattern Recognition, IEEE Computer Society Conference on, Vol. 1, pp. 886-893, 2005

[11]. Q. Xu and Y. Li, “Video classification using spatial-temporal features and PCA,” IEEE International Conference In Multimedia and Expo, Vol. 3, pp. 480-485, 2003

[12]. R. Nava, R. Boris, and C. Gabriel. "Texture image retrieval based on log-Gabor features." Springer Berlin Heidelberg in Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications, vol. pp. 414-421, 2012.

[13]. W. Qi, L. Gu, H. Jiang, X.-R. Chen, and H.-J. Zhang, “Integrating visual, audio and text analysis for news video,” IEEE International Conference on Image Processing, Vol. 3, pp. 520-523, 2001.

[14]. J. Nam, M. Alghoniemy, and A. H. Tewfik, “Audio-visual content-based violent scene characterization,” International Conference on Image Processing, Vol. 1, pp. 353–357, 1998

[15]. G. Tzanetakis, “Musical genre classification of audio signals”, IEEE Transactions on Speech and Audio Processing, Vol.10, pp. 293 - 302, 2002

[16]. G. Theodoros, D. Kosmopoulos, A. Aristidou, and S. Theodoridis. “Violence content classification using audio features”, Springer Berlin Heidelberg in Advances in Artificial Intelligence, Vol. 1, pp. 502-507, 2006

[17]. W. Qi, L. Gu, H. Jiang, X.-R. Chen, and H.-J. Zhang, “Integrating visual, audio and text analysis for news video,” IEEE International Conference on Image Processing, Vol. 3, pp. 520-523, 2001.