Volume: 3 Issue: II Month of publication: February 2015

DOI:

www.ijraset.com
Volume 3 Issue II, February 2015
ISSN: 2321-9653

International Journal for Research in Applied Science & Engineering Technology (IJRASET)

Video Genre Recognition Using Gestures of the Viewer

Prashant Dahiya[1], Samiulla Zakir Hussain Shaikh[2]
Computer Scientist, Adobe Systems India, Noida prashant.er@gmail.com
Member of Technical Staff, Adobe Systems India, Noida

Abstract: There have been many attempts of video genre recognition using the analysis of video itself, but none could perform close to the human intelligence. We propose an approach in which we will use viewers' reactions while watching the video for detecting its genre. Viewers hardly fail to identify the genre of the video, which in-turn is reflected through their gestures[1]. Hence the accuracy of the system will solely depend on the viewer's gesture recognition.

I. BACKGROUND

Majority of existing 'video genre detection' systems rely on analysis of various features of the video. Various successful approaches have used features like dynamics, camera motion, speech, music, color statistics etc. for genre recognition. Let's have a look at advantages and disadvantages of using above features.

A. Video dynamics (M.Pawlewski, 2001).
B. Camera motion, speech, object motion and audio (Stephan Fischer, 1995).

The above techniques are computationally intensive, cannot be used on real-time.

II. PRIOR ART/SOLUTIONS

A. "Apparatus and method for determining genre of multimedia data". (Doo Hwang, 2007)
They are using audio and video data to determine genre of the multimedia file. Their feature extraction is limited to audio and video data.
B. "Video genre categorization and representation using audio-visual information" (Ionescu, Seyerlehner, Rasche, & Lambert, 2012)
They have used the audio, color, temporal and contour extracted from the video for determining the genre. The accuracy is high, but the process of extracting features from the video is computationally intensive and hence time consuming.
C. "A semantic framework for video genre classification and event analysis" (You, 2010)
They have used semantic video analysis for genre detection. The process involves weakly supervised machine learning techniques. The contextual relationship of events and statistical characteristics of dominant events are used as features for training HMM, GMM etc.

III. DESCRIPTION

Our system will identify the genre of the given video based on the gestures of audience. We'll continuously monitor viewer's gestures using the camera. The identification of facial expressions and upper body gestures can be done in real time (Baltrusaitis, 2011). This sequence of gestures is associated with respect to its video frame, and weightage is given according to frame's position in time line of the video. This means, we will take the timing into consideration. For example: We have formulated algorithm in a way so that, it gives a gradual increase in weightage from beginning to end of the video. As the video approaches end, the importance of user's gestures and thereby its effect on genre detection will increase gradually. We have also considered the rate of change of gestures as a factor in genre detection. As we have analyzed, in cases where the gestures' weighted counts over a video length remains nearly same; their rate of change and the period of persistence does impact the overall genre. The system will have a database storing gestures of videos whose genres are known. This data set will be used to train the system. The trained system is further used for genre identification of new videos.

Please refer the Algorithm section for detailed procedure.

Once, the system has processed a video, it also identifies the "genres"

[1] By Gestures we mean only facial expressions and upper body gestures

IV. ALGORITHM

There are three major parts of the algorithm:

A. Creating Normalized feature vector for the given video.
B. Creating a Trained Model.
C. Genre Detection of new video.

All these are explained below:

A. Creating Normalized feature vector for the given video

Given a video as input; initially, the system will process it to create a unique unit vector. - First of all, we will need to process the video and users gestures, using a generic gesture recognition algorithm.

1) The output of this process will be continuously stored in an associative data structure such as map, to hold time spans with respect to "gestures". That means, gesture will be the key, and time spans will be values (M1) - <gesture,time span>.
2) The values in M1 will now be used to create the unit vector for the video.

$$f(t) = \frac{\log\left(\frac{\alpha.t}{T} + \beta\right)}{\log(\alpha + \beta)}$$

Equation 1 : weighing function for feature vector creation

Where,

T = total time of the video
α/T = needed to vary the slope of graph (m), so that "m" remains similar irrespective of the length of video
β = needed to maintain a non-zero (minimum) weight (at t = 0).

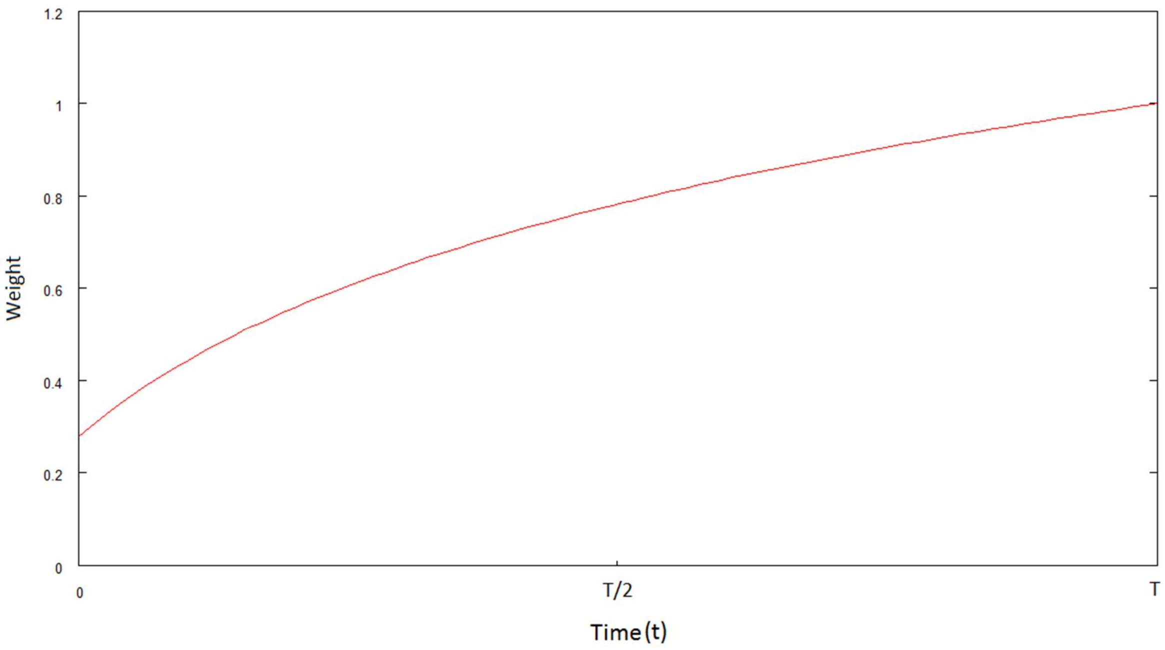

Figure 1: Weighing Curve depicting Equation 1 i.e. f(t)

$$\vec{V} = \left(\sum_{i\epsilon unit(gesture)} \sum_{t_{1j},t_{2j} \in timeRanges(gesture(i))} \left(\left(\int_{t_{1j}}^{t_{2j}} f(t).dt \right).\hat{\imath} \right) \right) + r.\hat{k}$$

Equation 2 : Equation to create feature vector for M1

Where,

i = unit vector representing each of the mutually orthogonal gestures (from M1)
j = index for the different time-ranges of the current gesture (i)
$[t_{1j},t_{2j}]$ = time interval for the Jth occurrence of the current gesture (i)
$\vec{V}$ = feature vector corresponding to the current video under process
r = average rate of change of gestures
$\hat{k}$ = unit vector along the feature "r"

Above equation will be used to form the feature vector for the video; based on the values captured in the map – M1.
For every gesture i, in M1 we take the weighted integral over each of the time ranges. The weighing factor "e^{t-T}" ensures that the later part of the video is given higher importance. Assumption here is that ending of a video has greater influence in genre determination. We have also considered rate of change of gestures as another dimension of feature vector.

As, we have recorded the relative-weighted frequency of gestures in above formed relation. We can now neglect the duration of video as a factor in genre detection. Thus, we can normalize the feature vector, in following way:

$$\hat{V} = \frac{\vec{V}}{|V|}$$

Equation 2 : Normalization of feature vector

B. Creating the Data Set

We will use the above mentioned algorithm to create a set of normalized feature vectors for the videos, whose genres are already known.
This model will be used for genre detection of unknown (test) videos.

$$\text{Data Set (D)} : \{\hat{V}_1, \hat{V}_2, \dots, \hat{V}_n\}$$

This is Data Set - D, which can directly be used by various supervised machine learning algorithms to learn the behavior of the data set.

C. Behavior learning from Data Set to create the Trained Model

This step/module depends on the algorithm used for classification. The output of this step (Trained Model – T) will be used along with the unit feature vector created for the test video in step 4.b to conclude the genre of the video.

For example:

1) Algorithm: K-Nearest neighbors : A set N will be created, which will contain the k-nearest neighbors of the test vector ($\hat{V}_t$) from data set - D.
2) Algorithm: Support Vector Machine : A set of binary Classifiers – C will be created, using vectors in the data set (from step 2).

D. Genre Detection of new video

Given a new (test) video; following steps are used to determine the genre by this system. This is the main Algorithm, which will use all the above mentioned modules. Note that, creation of Data Set and the corresponding Trained Model will be done only once using module -2 and 3.

For each test video (input), following steps will be performed for genre detection:

1) Create a normalized feature vector for the test video - $\hat{V}_t$. This can be done using the module 1 – "Creating Normalized feature vector for the given video".

The time required for this step is very less (Please see in Appendix A, for details).

2) Testing phase: Now, the trained model (T) will be used to compute scores for different genres with respect to the test vector - $\hat{V}_t$. The Genre with the maximum score will be assigned to the video.

The Testing phase takes nominal time.

Thus, the genre is identified in automated way taking the user's gestures in consideration.

V. THE USER'S WORKFLOW

The target user here are the organizations or people (we will call them "user" henceforth) involved in video creation and distribution. Every time consumer of the user is watching a video using their interface; the camera on consumer's device will be utilized (with consumer's permission) to capture gestures. These will be used to identify the genre of the video at run time.

VI. ADVANTAGES

The proposed solution is clearly more desirable than present techniques for genre detection because of vast improvements in time and space requirements. The time required is much less and thus this method can be used at real time, instead of separate processing. The advantage it has over Latent Semantic Analysis is that our system doesn't require a large corpus. Our system uses human intelligence to determine the genre; which is more accurate than the mechanisms or determining factors used in present techniques.

APPENDIX

A. Time calculations

The integral calculations take nominal time. The dependency is on the number of intervals (step-size). We are varying the step-size based on the slope of the curve, i.e. step-size will be inversely proportional to the slope. This improves the time required without compromising the accuracy. Time complexity for vector formation, is O(n). Where n is the number of intervals. This can be proved (calculated) using Gauss's rule.

REFERENCES

[1] Baltrusaitis, T. D. (2011). Real-time inference of mental states from facial expressions and upper body gestures. IEEE (pp. 909-914). IEEE.

[2] Doo Hwang, J. K. (2007). Patent No. US20070113248 A1. USA.

[3] Ionescu, B., Seyerlehner, K., Rasche, C., & Lambert, C. V. (2012). Video genre categorization and representation using audio-visual information. J. Electron. Imaging. 21(2), 023017 .

[4] M.Pawlewski, M. J. (2001). Video genre classification using dynamics. IEEE International Conference (pp. 1557-1560). Washington: IEEE Computer Society.

[5] Stephan Fischer, R. L. (1995). Automatic recognition of film genres. MULTIMEDIA '95 Proceedings of the third ACM international conference on Multimedia (pp. 295-304). New York: ACM.

[6] You, J. G. (2010). A semantic framework for video genre classification and event analysis. Signal Processing: Image Communication 25.4, 287-302.