Research Journal of Applied Sciences, Engineering and Technology 5(17): 4349-4355, 2013
DOI:10.19026/rjaset.5.4428
ISSN: 2040-7459; e-ISSN: 2040-7467

Submitted: August 21, 2012 Accepted: September 27, 2012 Published: May 01, 2013

Research Article

Multimodal Signal Comparison with the Aim of Video Content and Quality Monitoring for IPTV Application

[1, 2]Jila Hosseinkhani, [1]Hassan Yeganeh and [1]Mehdi Samie
[1]Iraninan Telecommunication Research Center (ITRC), Tehran, Iran
[2]CIPCE, Department of Electrical and Computer Engineering University of Tehran, Tehran, Iran

Abstract: In the present study the video content and quality monitoring issue is studied. For this purpose, the contents and qualities of two video streams must be compared with each other. It is clear that achieving to a monitoring system will be possible by utilizing multimodal information from the video streams. Therefore, simultaneous evaluation of the image and audio signals is vital. This comparative study for image signal is based on the extraction of useful features such as texture using Gabor filter. In order to create differences between two video streams in terms of content, some frames are added to or eliminated from the original video. Moreover, distortions such as blurring, packet loss and adding three types of noises are done as attacks to the original videos in order to make difference in terms of quality. Three types of additive noises such as Gaussian, Poisson and Speckel are used to produce noisy images for the purpose of comparing with the original ones. The audio signals of the two compared video streams are evaluated using PESQ similarity measurement. Finally, these parameters are characterized on the basis of some statistical standard image quality matrices like SNR, Correlation coefficient and SSIM. The results illustrate that the proposed method is effective and highly reliable against various kinds of noises.

Keywords: Gabor filter, PESQ, SSIM, texture extraction, video monitoring

INTRODUCTION

In recent decades, video analysis has gained considerable attention in computer vision and image processing societies. Content-based video analysis is indispensable in order to effectively manage and utilize these kinds of data.

Consistent with the rapid growth of imaging on the World Wide Web and broadcasting requires video monitoring system, which can filter unexpected content from the video streams, detect and recognize images, video and multimedia data.

Because of publicity and fragility problems which video broadcasting is faced, it is vulnerable in front of attacks. For example, some of the fundamental issues in this area include the change of video content, the insertion of illegal video, the change of audio content/ energy/ frequency and the replacement of unauthorized information (Liu and Zhan, 2008).

The first step to compare and match two video streams is feature extraction from both of them. Then the selected features are compared with each other using a group of metrics. Obviously, selecting the appropriate features is highly important and depends on the application.

It should be noted that texture is an important component in perception, classification, identification and segmentation of images. There are many techniques have been used for measuring texture similarity.

In Wang *et al*. (2006) a camera motion estimation algorithm is utilized to detect visual signal changes. Indeed, it is focused on the association of visual signal changes (e.g., cuts, fade-in, fade-out, etc.) and audio signal changes (e.g., speaker change, background music change, etc.). The main limitation and weakness of this approach is that the visual signal change may not synchronize with audio signal change; therefore it causes difficulties for scene change detection process.

In Kenyon and Simkins (1991) the problem of real time monitoring of broadcast radio and television stations is considered. Indeed, the main purpose is to identify when and where specific musical recordings and advertisements are transmitted. The proposed method involves a two stages pattern recognition procedure to eliminate unlikely candidate signatures.

In addition, some works consider the protection of copyrighted video data. For example, in Yoshida and Murabayashi (2008) a content-based approach has been proposed which utilizes RGB-based video signature and hash functions to detect copied video. The copied video is detected by comparing the uploaded files with stored video data. Obviously, alteration of the video, such as rescaling and cropping, makes this task more difficult.

In Erol and Kossentini (2001) color histogram of frames is used for content detection during the video

Corresponding Author: Jila Hosseinkhani, Iraninan Telecommunication Research Center (ITRC), Tehran, Iran

streams. According to real-time system requirements, reduction of the computational variables and cost is essential. Given this, the developed method in Pribula *et al*. (2010) is based on the brightness profile processing of significant image regions. Algorithm output is a one dimensional discredited time vector of video sequence fingerprints that encompasses both temporal and spatial video content information. Since the size of these fingerprints is significantly lower than the source video, their storing in database is very appropriate.

In Neto *et al*. (2011) an automatic monitoring system is developed and built for TV and radio channels information. The proposed approach is based on the speech recognition techniques in order to monitor video sequences. The main focus is on the annotation of broadcast news, where its content is being transcribed in the video. But transcribing a video can be a challenging task due to acoustic condition, speaker variability and information contents. On the other hand, major efforts were made in this area mainly for English in the last decades. However, not much work has been done for other languages.

One of the useful ways for preserving the security of multimedia information is using digital watermarking technologies. Digital watermarks are embedded in broadcast video streams as impalpable codes. They can be extracted at receivers using relevant algorithms by special hardware at the control stations. This makes possible to exactly follow the video content when it is broadcasted. Moreover, hidden codes detection can be done as a real time process. For example, in Ong *et al*. (2009) MPEG-2 bit stream is used for video data watermarking. However, the most important limitation of utilization of watermarks is that the original data mt be available at the transmitter side.

There are a variety of measurements to evaluate streaming video quality, so that the large numbers of them are based on the perception. Some of quality assessment techniques utilize the original video stream as a reference and some others require no reference for evaluation process. In addition, there is an eminent requirement for in-service analysis and quality measurement of streaming video where the original undistorted videos are not available for comparison. Therefore, the use of no-reference techniques becomes more significant in the multimedia industry. In Clausi and Deng (2005) a no-reference and real time method is described.

Generally, in this study, two video streams are compared with each other and the proposed method can distinguish whether two streams are the same video or not. If two videos are the same then our proposed method can compare them from quality aspect. In this way, we are able to recognize the probable noises and artifacts which are created during data transferring over communication channels and networks. It should be noted that, the proposed method can be utilized as a part of the Internet Protocol TV (IPTV) and Video on Demand (VOD) equipments for video monitoring, video retrieval and video filtering purposes.

The main characteristic of our proposed method for video streams comparison is the high reliability against noises and artifacts and its ability in discrimination of two different video streams.

ALGORITHM DESCRIPTION

Overview: As mentioned before, there exist an extension need of video content matching and monitoring algorithm with efficient and sufficient performance which can be easily and fast implemented with low computational cost.

The proposed algorithm uses texture features to match two video sequences. Since the algorithm must be reliable against different changes and attacks, at first some changes are applied over the original video data. Then the changed video stream is compared with the asset video stream to find probable mismatches from both content and quality aspects. However, comparison stage is highly depending on the appropriate and effective selected features which are extracted from both of original and streamed videos. The comparison includes two main stages. Firstly, it should be found out that two video streams are the same or not. Secondly, they are evaluated using statistical similarity metrics such as Correlation coefficient and SSIM to realize noisy frames.

The proposed framework has three main steps which are described by details in the next sections.

Separating audio and video signals: At first it is necessary to separate the audio and video signals from each other. According to the performed investigation, it is concluded that AoA Audio Extractor software can be the appropriate choice among the various available pieces of software in this area. After extraction audio from the video, both types of signals should be processed by the proposed algorithms.

Applying changes over the video frames: Obviously, before comparing two video sequences, their content and quality must be changed and attacked by various noises. Therefore, the used way for changing the content of video sequence is adding a series of frames amidst the original video frames or reducing some frames from its own frames. In this way, video content undergoes changes which make it different from its original status. Moreover, some other changes are applied to the original video altering its quality as well as the content. For example, resolution and frame ratio are the scales that can be changed for this purpose. Besides, three types of additive noises such as Gaussian, Poisson, Speckel are used to produce noisy images.

Evaluating video by texture extraction using Gabor filter: Feature extraction is the most critical phase in the video comparing and matching process. Two proper and useful features for this goal are color histogram and texture. It is worthwhile to mention that texture is more efficient than color histogram because the image color is influenced by lighting condition while the video is recording and even by the amount of display screen light.

Therefore, we selected texture as a prominent feature to compare video streams and detect probable changes.

Generally, there are so many approaches to extract the video content which the most significant of them are object detection, object tracking during consecutive frames, shot segmentation and texture segmentation.

Since the characters that play role during the film are numerous, so it is not reasonable to use methods like object detection and tracking. This leads to the high computational cost and wastes the time. On the other hand, the existent characters in a film are absent in some frames and appears in some others. Consequently, it will not be able to identify the occurred changes effectively for each frame. In contrast, extracting various textures in an image is a challenging problem but very beneficent and helpful information can be obtained using them. Texture segmentation is the task of identifying regions with similar patterns in an image. There is no known method that is able to consistently and accurately segment textured images. A commonly used strategy for this purpose is firstly to extract features on a pixel-by-pixel basis from an image and then use some technique to classify the extracted features (Clausi, 2002; Randen and Husoy, 1999; Jain and Farrokhnia, 1991; Marcelja, 1980). Generally, two popular methods for texture feature extraction are grey level co-occurrence probabilities (GLCPs) and Gabor filters.

The main motivation to use Gabor filters is that receptive fields of simple cells in the primary visual cortex of mammals are oriented and have characteristic spatial frequencies (Andrysiak and Chora's, 2005). These could be modeled as complex 2-D Gabor filters (Movellan, 2002). The limitation of GLCPs method is that it works very well in high frequencies so; its performance is influenced by this condition. Generally, Gabor filter bank is a good choice in comparison with other methods to extract image texture because of its high accuracy and high adaptation with human visual system. Although, this method is not so fast, but as regards designing a real time algorithm was not our purpose so it cannot make a challenge and problem. Gabor filters will be investigated in details afterwards.

Introduction of Gabor filter: A Gabor filter is a linear filter whose impulse response is defined by a harmonic function multiplied by a Gaussian function. They are directly related to Gabor wavelets, since they can be designed for a number of dilations and rotations. Therefore, usually a filter bank consisting of Gabor filters with various scales and rotations is created. The filters are convolved with the signal, resulting in a so-called Gabor space. This process is closely related to processes in the primary visual cortex. The Gabor Filters have received considerable attention because the characteristics of certain cells in the visual cortex of some mammals can be approximated by these filters.

In addition, these filters have been shown to posses optimal localization properties in both spatial and frequency domain and thus are well suited for texture segmentation problems.

Gabor filters have been used in many applications, such as texture segmentation, target detection, fractal dimension management, document analysis, edge detection, retina identification and image coding and image representation. A Gabor filter can be viewed as a sinusoidal plane of particular frequency and orientation, modulated by a Gaussian envelope (Abhay, 2009).

Here is the formula of a complex Gabor function in space domain:

$$g(x, y) = s(x, y)w_r(x, y) \tag{1}$$

where, $s(x, y)$ is a complex sinusoid, known as the carrier and $w_r(x, y)$ is a 2-D Gaussian shaped function, known as the envelope.

Also, the complex sinusoid is defined as follows:

$$s(x, y) = \exp(j(2\pi(u_0 x + v_0 y) + P)) \tag{2}$$

where, (u, v_0) and P define the spatial frequency and the phase of the sinusoid respectively. The Gaussian envelope also looks as follows:

$$w_r(x, y) = K \exp(-\pi(a^2(x - x_0)_r^{\ 2} + b^2(y - y_0)_r^{\ 2}) \tag{3}$$

where, (x_0, y_0) is the peak of the function, a and b are scaling parameters of the Gaussian and the r subscript stands for a rotation operation e.g., θ:

$$(x - x_0)_r = (x - x_0)\cos\theta + (y - y_0)\sin\theta \tag{4}$$

$$(y - y_0)_r = (x - x_0)\sin\theta + (y - y_0)\cos\theta \tag{5}$$

Description of similarity metrics: After extracting the texture of two video streams, it is essential to obtain the correlation coefficients between two corresponding textured frames.

Correlation operation is closely related to convolution. In correlation, the value of an output pixel is calculated as a weighted sum of neighboring pixels. Correlation coefficient matrix represents the normalized measure of the strength of linear relationship between variable correlation coefficients. Indeed, correlation

denotes the strength and direction of the linear relationship between two signals and its value lie in interval [-1, 1]. However, the values closer to either +1 or -1 indicate the stronger linear relationship between signals and some value in between for all other cases, including the degree of linear dependence between the two signals.

The Eq. (6) denotes the normalized correlation coefficients between two signals:

$$COC = \frac{\sum (g - \bar{g})(\hat{g} - \bar{g})}{\sqrt{\sum (g - \bar{g})^2 \sum (\hat{g} - \hat{\bar{g}})^2}} \quad (6)$$

where, g and $\hat{g}$ are the original and changed frames and also $\bar{g}$ and $\hat{\bar{g}}$ are the means of the original and the changed frames respectively.

Noise recognizing using SSIM statistical metric: Once it is proved that two video streams belong to a single video and then the presence of noise must be recognized. For this purpose, SSIM (Structural Similarity) is an applied metric for comparison between the original and disturbed signals. This index is a method for measuring the similarity between two images. The SSIM index is a full reference metric, in other words, the measuring of image quality based on an initial uncompressed or distortion-free image as reference. SSIM is designed to improve on traditional methods like Peak Signal to Noise Ratio (PSNR) and Mean Squared Error (MSE), which have proved to be inconsistent with human eye perception. This metric is calculated on various windows of an image. The measure between two windows x and y with the size of N×N is as follows:

$$SSIM(x, y) = \frac{(2\mu_X \mu_Y + c_1)(2\,\mathrm{cov}_{xy} + c_2)}{(\mu_x^{\,2} + \mu_y^{\,2} + c_1)(\sigma_x^{\,2} + \sigma_y^{\,2} + c^2)} \quad (7)$$

In this equation, μ_x and μ_y respectively indicate the mean value within windows x and y. The covariance value of x and y is showed as cov_{xy}. Also, c_1 and c_2 are two variables which cause to stabilize the division with weak denominator and are defined as follows:

$$c_1 = (k_1 L)^2 \quad (8)$$

$$c_2 = (k_2 L)^2 \quad (9)$$

where, L is the dynamic range of pixel values. $k_1 = 0.01$ and $k_2 = 0.03$ by default.

Therefore, the noisy and destroyed frames can be recognized accurately using SSIM index (Klatt, 1982).

PESQ measurement: Generally, there are two main methods for evaluating speech quality including subjective and objective methods. However, the most accurate method for this purpose is through subjective listening tests. Although, subjective evaluation of speech enhancement algorithms is often accurate and reliable, but it has restrictions like being costly and time consuming. For that reason, it has been tried on developing objective measures that would predict speech quality with high correlation (Hu and Loizou, 2008). Among all objective measures, the PESQ (Perceptual Evaluation Speech Quality) measure is the most complex to compute and is the one recommended by ITU-T for the aim of speech quality assessment.

The PESQ score is defined as a linear combination of the average disturbance value D_{ind} and the average asymmetrical disturbance values A_{ind} as follows:

$$PESQ = a_0 + a_1 D_{ind} + a_2 A_{ind} \quad (10)$$

where, $a_0 = 4.5$, $a_1 = -0.1$, $a_2 = -0.0309$, and . These parameters are optimized for speech processed through networks and not for speech enhanced by noise suppression algorithms.

It is necessary to mention that, in the PESQ measure the original and degraded signals are mapped onto an internal representation using a perceptual model. The difference in this representation is used by a cognitive model to predict the perceived speech quality of the degraded signal. This perceived listening quality is expressed in terms of Mean Opinion Score (MOS), an average quality score over a large set of subjects.

EXPERIMENTAL RESULTS

The proposed method has been tested on two video streams and the algorithm has been implemented using MATLAB.

The suggested method for video content and quality comparison was tested on 145 frames of Video1 and on 55 frames of Video2.

The important properties of the video sequences which were used to test our method are as follows: Video1 and Video2 have 144×176 and 720×1280 resolutions respectively and both of them have 25 frame rates.

In order to destroy the video signal quality, we added zero mean Gaussian noise with the variance value of 0.001, Poisson noise, speckel noise with the variance value of 0.01 and also applied some artifacts like packet loss and blurring.

Table 1 shows the results of content comparison by texture extraction using Gabor filter bank. In this table, the first row indicates the type of artifacts and noises,

Table 1: Results of content comparison by texture extraction

Sample	Video1		Video2		Video1		
Noise type	Blurring	Gaussian noise	Original video	Position and packet loss	Packet loss	Speckel	Original video
Frame number	35	35	40	45	30	30	25
Correlation	0.9637	0.9975	0.3785	0.4580	0.8666	0.9610	0.9713
SNR-noisy (db)	16.9833	14.2387	-3.0747	-2.6998	14.5170	14.2150	14.9396
SNR-Filtered (db)	10.7768	13.5772	-4.5661	-4.4306	14.6592	12.8613	15.9682

Table 2: Results of quality comparison using ssim index

Sample	Video1		Video2	Video1	
Noise type	Blurring	Gaussian	Poisson and packet loss	Packet loss	Speckel
Frame number	35	35	45	30	30
SSIM	0.4783	0.8080	0.1482	0.7084	0.6883

(a) (b) (c)

Fig. 1: Textured frames from video1 and video2

(a) (b) (c)

Fig. 2: The destroyed frames of video1 and video2

the second row shows the number of tested frames and finally the third row indicates the mean value of correlation coefficients between two textured video streams.

Based on experimental results, it can be concluded the proposed method is very effective and highly reliable against noise and artifacts. Texture is a strong feature is able to discriminate between two different contents.

Table 2 shows the results of the quality evaluation process due to SSIM metric. As in the previous table, the first and the second rows indicate the type of artifacts and noises and the number of tested frames respectively. Numbers in the third row show the mean value of SSIM calculated for noise realization.

In addition, Fig. 1 shows samples of extracted texture information. In this Fig. 1a and b show the textured frames from Video1 and Fig. 1c shows this information for a frame of Video2.

Figure 2 shows the corresponding frames with the textured ones in Fig. 1.

However the human visual system contains many filters with different peak spatial frequencies operating in parallel and it is reasonable to assume that discrimination is based on the filter that yield the best discrimination for the pair of patterns involved.

The various scales of PESQ measurement are interpreted in Table 3. According to the information of this table, if the PESQ measure is more than 3 it can be

Table 3: Quality of the speech score

Excellent	5
Good	4
Fair	3
Poor	2
Bad	1

Table 4: Composition objective measure

LRR	1.029096
SNRseg	10.628516
WSS	50.469029
PESQ	2.776189
Covl	3.2539
Csig	3.2773
Cbak	2.9487

concluded that the original and changed signals are very similar and noises are not so noticeable.

In this study, The Weighted Spectral Slope (WSS) function implements the composite objective measure proposed in Randen and Husoy (1999). It returns three values: The predicted rating of overall quality (Covl), the rating of speech distortion (Csig) and the rating of background distortion (Cbak). The ratings are based on the 1-5 MOS scale. In addition, it returns the values of the SNRseg, Log-Likelihood Ratio (LLR), PESQ and Weighted Spectral Slope (WSS) objective measures. According to the assumptions and results of Hu and Loizou (2008) three rates Csig, Cbak and Covl are respectively calculated by the below equations:

$$Csig = 3.093 - 1.029 \times LLR + 0.603 \times pesq - 0.009 \times WSS \quad (11)$$

$$Cbak = 1.634 + 0.478 \times pesq\ 0.007 \times WSS + 0.063 \times SNRseg \quad (12)$$

$$Covl = 1.594 + 0.805 \times pesq - 0.512 \times LLR - 0.007 \times WSS \quad (13)$$

Table 4 shows the results of calculated composite objective measures. The PESQ measure illustrates that comparison between original voice signal and the noisy one has quite reasonable result.

CONCLUSION

In this study, a new method for video content and quality comparison is introduced. This method can be used in order to monitor video signals for IPTV and VOD (video on demand) applications.

In this study, a combination of texture features and quality measurements like correlation and SSIM index are used to realize and identify the changed or destroyed video streams. In addition, the PESQ measurement is used to do comparison between the original and destroyed audio signals. Clearly, utilizing multimodal features will be very effective and helpful for comparing two video streams.

This study can be utilized as a part of the IPTV and VOD equipments for video monitoring, video retrieval and video filtering.

Our future study will include multimodal signal processing for this purpose. In the other words, the audio signal can play important role in quality evaluation for video streams. Therefore, usage of this kind of information will be highly effective in the proposed algorithm performance.

REFERENCES

Abhay, K., 2009. Evaluation of Gabor filter parameters for image enhancement and segmentation. M.Sc. Thesis, Department of Engineering, Thapar University, Punjab.

Andrysiak, T. and M. Chora's, 2005. Image retrieval based on hierarchical Gabor filters. Int. J. Appl. Math. Comput. Sci., 15(4): 471-480.

Clausi, D.A., 2002. K-means Iterative Fisher (KIF) unsupervised clustering algorithm applied to image texture segmentation. Pattern Recognit., 35(9): 1959-1972.

Clausi, D.A. and H. Deng, 2005. Design-based texture feature fusion using Gabor filters and co-occurrence probabilities. IEEE T. Image Process., 14(7): 925- 936.

Erol, B. and F. Kossentini, 2001. Color content matching of MPEG-4 video objects. IEEE Pacific Rim Conference on Multimedia, Beijing, China, pp: 891-896.

Hu, Y. and P.C. Loizou, 2008. Evaluation of objective quality measures for speech enhancement. IEEE Trans. Audio, Speech Language Proces., 16(1): 229-238.

Jain, A.K. and F. Farrokhnia, 1991. Unsupervised texture segmentation using Gabor filters. Pattern Recognit., 24(12): 1167-1186.

Kenyon, S.C. and L. Simkins, 1991. High capacity real time broadcast monitoring: Systems, man and cybernetics, 1991. Proceedings of IEEE International Conference on Decision Aiding for Complex Systems, Charlottesville, VA, 1: 147-152.

Klatt, D., 1982. Prediction of perceived phonetic distance from critical band spectra. Proceeding of IEEE International Conference on Acoustics, Speech and Signal Processing, 7: 1278-1281.

Liu, H. and D. Zhan, 2008. A content monitoring system based on MPEG-2 video watermarking. International Symposium on Intelligent Signal Processing and Communication Systems ISPACS, Xiamen, pp: 546-549.

Marcelja, S., 1980. Mathematical description of the responses of simple cortical cells. J. Opt. Soc. Am., 70: 1297-1300.

Movellan, J.R., 2002. Tutorial on Gabor filters. Technical Report.

Neto, J., H. Meinedo and M. Viveiros, 2011. A media monitoring solution. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Prague, pp: 1813-1816

Ong, E.P., S. Wu, M.H. Loke, S. Rahardja, J. Tay, C. Tan and L. Huang, 2009. Video quality monitoring of streamed videos. IEEE International Conference on Acoustics, Speech and Signal Processing ICASSP, Taipei, pp: 1153-1156.

Pribula, O., J. Pohanka and J. Fischer, 2010. Real-time video sequence matching using the spatio-temporal fingerprint. 15th IEEE Mediterranean Electro-technical Conference MELECON, Valletta, pp: 911-916.

Randen, T. and J.H. Husoy, 1999. Filtering for texture classification: A comparative study. IEEE T. Pattern Anal., 21(4): 291-310.

Wang, J., L. Duan, H. Lu, J.S. Jin and C. Xu, 2006. A Mid-Level scene change representation via audiovisual alignment. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp: 409-412.

Yoshida, K. and N. Murabayashi, 2008. Tiny LSH for content-based copied video detection. International Symposium on Applications and the Internet (SAINT), Turku, pp: 89-95.