International Journal of Engineering Research and Technology. ISSN 0974-3154 Volume 10, Number 1 (2017)

Temporal Dynamics of Continuous Facial Emotion Recognition System

Anil S. Patthe
Research Scholar, Department Of Computer Science and Engineering,
Walchand College of Engineering,
Sangli, Maharashtra, India, 416415
pattheanil@gmail.com

Mr. Anil R. Surve
Assistant Professor, Department Computer Science and Engineering,
Walchand College of Engineering,
Sangli, Maharashtra, India, 416415.
anil.surve@walchandsangli.ac.in

Abstract
We are now in new era of technology where even human cannot understand the emotion of another human. In such cases computers can be used to detect human emotions. But this too has a limitation as sometimes it's difficult for the algorithm to detect real emotions from the fake ones[2]. Growing need for emotion recognition in today's fast growing technological world, this field needs research contribution. This paper proposed a system we propose a system to make existing emotion recognition more efficient. High resolution and higher megapixel cameras are used to collect emotion related data from real world. To minimize the storage requirement of such large data efficient methodology is proposed[1].

Keywords: Affective computing, SVM, sampling, Discrete Fourier Transform (DFT)

Introduction
Facial expression recognition is so far the best way of communication between humans. Thus it can be used to automatic systems to eliminate the human intervention. Such system can be used in Human Computer applications e.g. baby monitoring, Driver safety, Health Care, advertising etc. this has led to huge demand for affective computing in the field of computer vision for face detection and recognition[2]. Affective computing deals with collecting human emotions and keep relevant information, elimination the need of storing redundant data. The processed data then can be used to analyze and make conclusions about human behavior. Thus the helping users to perform their actions as required.

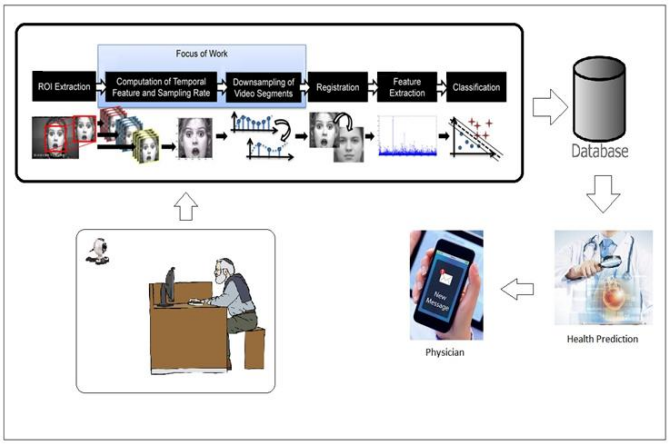

Fig. 1

Such automated system can be very useful in the field of Internet of Things, as it involves handling large amount of data simultaneously. Any centralized IoT system consist number of cameras which collect images everyday generating large database of image frames. To extract relevant emotion frames from such database is tedious task. The proposed system does the work of eliminating redundant frames dynamically from the system. The important frames are then transferred over IoT network. This helps reduce the load of IoT network in terms of utilization of bandwidth, storage and processing power making system efficient.

Methods
For long period of time the high frame rate videos are captured continuously[1]. In such cases it becomes difficult to model and train classification for all the frames present in given dataset. One way to get over this is uniform temporal downsampling of the video at low frame rate. This approach fails to detect precisely the changes in emotions and results in loss of precision. Dynamically assigning frame rate resolves the issue i.e. Dynamic sampling rate[1]; for videos involving idle or no motion lower frame rate is assigned and for higher

frame rate is assigned for person in continuous motion.

In proposed method we are going to minimize the extra overhead of system so we can recognize the emotions in the efficient time, storage and processing dynamically. However, the work focus is the reduction of frames which is not able to convey the emotions and which having the redundant information in sequential frames. For a part of a scene, the amount of attention given is affected by the temporal frequency of visual frequency in the scene.

The overview of this work is shown in Fig. 2:

1) Frames are captured in real time scenario by using high resolution cameras. These frames are transferred to local processing unit for frames preprocess.
2) At local preprocess system frames are converted from RGB to Gray scale images.
3) From each and every frame faces ROI (Regions of Interest)[10] are detected by using Viola-Jones and KLT algorithm.
4) The collection of all frames is partitioned into further segments. In every segment, the quantification of visual information is performed with temporal frames.
(5) Then all frames are downsampled to the dominant frequency.
(6) After that by using avatar the best image registration method we are going to align those frames.
(7) For all the selected frames in local region appearance features are generated
(8) All selected downsampled frames will be the input for the emotion labeling method after the classification of each frame into the particular emotion class by using the SVM[12] classification algorithm.

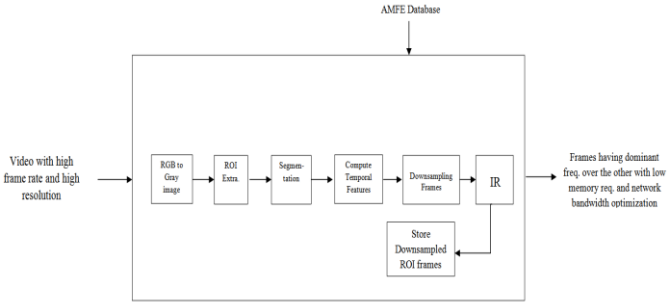

Fig. 2

Downsampling Continuous Video: As the data enters the system, the continuous video is downsampled without annotations for apexes. These videos are then uniformly split into smaller segments. Every segment with its own downsampling factor is downsampled dynamically. This is not the case in normal scenarios as they involve uniform downsampling rate. Downsampling method is explained in algorithm below.

Algorithm: Computing the sampling rate for single segment.

Input: I_Φ, the video segment. n_0, midpoint-apex time point (if given). N, number of frames in Φ.

Output: I_{Φ^*}, downsampled video segment.

procedure DOWNSAMPLESEGMENTS(I_Φ)
  for all frames $n \in \Phi$ **do**
    $\Delta I_n \leftarrow$ optical flow from n -1 to n
    $f(n) = \sum_{\mathbf{x}} \| \Delta I_n (\mathrm{x}) \|_2$
  end for
  $\mathbf{f}_\Phi \leftarrow$ vector corresponding to all features f
  $\tilde{\mathbf{f}}_\Phi \leftarrow \mathbf{f}_\Phi$ - mean of $\mathbf{f}_\Phi$
  $\mathbf{F}_\Phi \leftarrow$ Discrete Fourier transform of $\tilde{\mathbf{f}}_\Phi$
  $\beta \leftarrow \mathrm{argmax}_k \| \mathbf{F}_\Phi (k) \|$
  if n_0 is given **then**
    $\Phi^*_{\mathrm{Apex}} \leftarrow$ range $n_0 - \beta/2 < n \leq n_0 + \beta/2$
    $\Phi^* \leftarrow \Phi^*_{\mathrm{Apex}}$
  else
    $M \leftarrow N/\beta$ ....(Downsampling factor)
    $\Phi^* \leftarrow \Phi \downarrow M$ ....(Every M-frame)
  end if
  return I_{Φ^*}
end procedure

1st Phase: Procedure for Time Partitioning:
The captured image set I is further segmented into segments of equal sizes with N frames each. For segment $I\Phi$, frames are present at indices Φ where $\Phi = \{m_0, m_{0+1}, \ldots\ldots, m_{0+N-1}\}$. $I\Phi^*$ denotes a downsampled video segment with frames at indices Φ^*, where Φ^* denotes order of subset of Φ. In the beginning the system is delayed for N frames, this video segment with N frames at a time is processed first. Starting with $m_0 = 0$ indicated first N frames then $m_0 = N$ indicating frames from N to $2N-1$, such segments are created for entire video. For remaining, if any such frames form its own segment. According to vision theory the HVS allows maximum bound of 1Hz which restricts the value of N parameter to be 1 second for every segment.

2nd Phase: In second phase of downsampling $I\Phi$ is re-sampled at lower frequency to compute the temporal feature of $I\Phi^*$. Initially the facial expressions are quantified into a signals varying with time. The changes in facial expression are supposed to reflect in signal's frequency. ROI being a frontal face and high frame rate, facial expression are quantified by exploiting optical flow. For frames ranging from I_n to I_{n-1}, optical flow is represented by ΔI_n. It computes a motion vector. 1D signal is formed by summing up the magnitude for all pixels present in the image.

$$f(n) = \sum_{\mathbf{x}} \| \Delta I_n (\mathrm{x}) \|_2$$

For a single frame, $f(n)$ represents its temporal feature. x represents a pixel, and $\|.\|_2$ as the magnitude. The temporal feature f_Φ for a segment $I\Phi$ is given as:

$$\mathbf{f}_\Phi \equiv [\, f(m_0) , f(m_0 + 1), \ldots\ldots, f(m_0 + N - 1)]$$

Segmentation of video, computation of optical flow and generation of temporal flow is show in Fig. 2. Optical flow is computed before registration as the registration process is costly, thus resulting in registration of less number of frames. Optical flow is neither used for alignment nor classification.

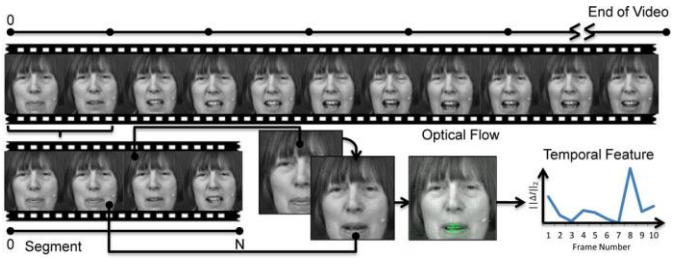

Fig. 3

3rd Phase: Video Segments downsampling

In next phase we have to compute the dominant frequency of one frame over the other frames. This dominant frequency we can compute by removing the DC-offset. We remove the DC-offset by:

$$\tilde{f}_{\Phi} = f_{\Phi} - E(f_{\Phi})$$

In above expression $E(.)$ is the value operator. There is important to compute the dominant frequency for the better result of downsampling so we remove the DC-offset here is two reasons for the renovation of DC-offset: (1) by removing DC-offset we can normalizes the given temporal feature and (2) second reason is for achieve the real data through the raw captured data, coefficient at 0Hz the $F_{\Phi}(0)$ corresponds, DC-offset may be higher than the values of F_{Φ}, so that we can get the dominant frequency. F_{Φ} is the DFT of $\tilde{f}_{\Phi}$: F_{Φ} = DFT ($\tilde{f}_{\Phi}$), Frequency-index k corresponds to the frequency with the most energy β which is calculated as follows:

$$\beta = \text{argmax} \, \| F_{\Phi}(k) \|,$$

In above expression k denotes frequency index, $\| F_{\Phi}(k) \|$ is the magnitude of $F_{\Phi}(k)$. We have to keep in mind that in above expression frequency isn't Nyquist-rate. For sampling a continuous signal in accurate reconstruction signal we apply Nyquist rate. We remove samples from the signal which is not having many changes with respect to the other samples. So we achieve this by computing its dominant frequency.

M is the downsampling factor and it is calculated by: Maximum frequency / Dominant frequency. β is the frequency index which can be replace to the dominant frequency by: $2\pi\beta/N$. Frequency 2π is corresponds by max frequency indexes N. It follows that: $M = N/\beta$. If $\Phi^* = \Phi \downarrow M$. So that Φ^* is the every M^{th} frame of Φ. At the time of high frequency of temporal features, $\beta \rightarrow N$, downsampling-factor will near about 1, and that time we save the all frames of that. But we remove the all frames at the time of downsampling-factor increase. It happed by the low frequency of temporal features.

Results

When we apply this system over the video segments the frames are downsampled dynamically so we can achieve the goal. In following example there are 7 frames in the video segment I with expression intensity. When we uses the random downsampling method over video segment I some weak frames having the low expression intensity are selected which is not convey the more information about the emotion.

Fig. 4

And when we apply the dynamic downsampling method over the same video segment only few frames are selected which having the high expression intensity over the other frames. That selected frames are more useful rather than the frames which are selected in random downsampling method.

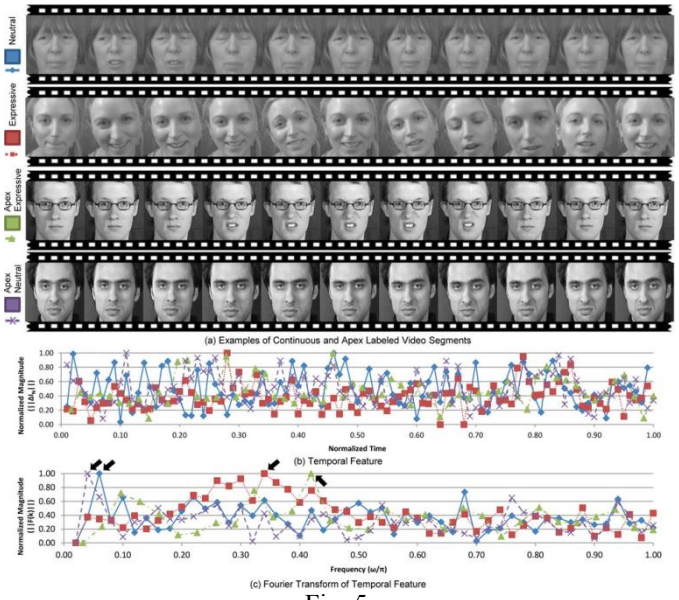

Fig. 5

Now we are discussing the cost saving of memory for dataset in following section and also the example for the temporal features: In the AMFED dataset, there are 242 total numbers of videos and total 168359 frames for the training, testing and development. Through the proposed downsampled method we downsampled the no. of frames by its factor 16.60 given in Table 1.

	AVEC 2011	AVEC 2012	CK	MMI-DB
# of Videos	74	95	488	222
# of Frames	1090476	1351129	8795	23466
Proposed	65871	76960	**1536**	**764**
Dahmane [8]	196051	239920	-	-
Savran [21]	-	232600	-	-
Glodek [13]	**740**	**950**	4930	2220

Table 1

It returns the outliers on the basis of regression labeling so that this is applied only over intensities of continuous labels. In this method every test frame is processed uniformly. All continuous data is classified into the 6 types of labels. We can be used proposed method for reduction of the number of frames from the all datasets. Proposed method can reduced the frames over the discrete data, continuous data, un-segmented data or segmented data.

Discussion

Based on the all results of a comprehensive assessment it was found that the continuous facial emotion recognition system using temporal dynamics[1] in nature gives classification of emotion in fraction of time. Also provide the appropriate emotion label to every frame. This system will work on the small amount of physical data storage and low processing power hardware. Only the limitation of this system is assumption where we assume that SVM algorithm labeling every frame correctly in respective emotion class.

Conclusion

This paper has presented efficient facial expression recognition system for accurate and better use of limited resources like low data storage hardware, low bandwidth network and low processing power hardware in IoT.

By using Continuous Facial Emotion Recognition System for Ambient Living we can predict and minimize health related problem which has possibilities to come in future. We can also use this system in e-learning for recognizing student's currant emotional state. And by temporarily downsampling the video frames we can minimize TB of Space which required to storing the all videos for future use. Also we efficiently recognized the user emotion in real time domain.

References

[1] Albert C. Cruz, Bir Bhanu, Ninad S. Thakoor. "Vision and Attention Theory Based Sampling for Continuous Facial Emotion Recognition", IEEE Transactions On Affective Computing, VOL. 5, NO. 4, October-December 2014.

[2] Mizna Rehman, Mamta Bachani, and Sundas Memon. "Blue Eyes Technology" Mizna m.rehman13@live.com, , 2013.

[3] S.Madhumitha, Slide Share, Blue Eyes Technology, March2013, <www.slideshare.net /Colloquium /blue-eyes-technology>.

[4] Daniel Hoglind, "Texture Based Expression Modeling for a Virtual Talking Head", Master of Science Thesis Stockholm Sweden 2006, < www.cse.kth.se >.

[5] B. Jiang, M. Valstar, B. Martinez, and M. Pantic, “A dynamic appearance descriptor approach to facial actions temporal modeling”, IEEE Trans. Syst., Man, Cybernetics Part B, vol. 44, no. 2,pp. 161–174, 2013.

[6] A. Savran, H. Cao, M. Shah, A. Nenkova, and R. Verma, “Combining video, audio and lexical indicators of affect in spontaneous conversation via particle filtering”, in Proc. 14th ACM Int. Conf. Multimodal Interaction Workshops, 2012, pp. 485–492.

[7] B. Schuller, M. F. Valstar, F. Eyben, R. Cowie, and M. Pantic,“AVEC 2012 the continuous audio/visual emotion challenge”, in Proc. 14th ACM Int. Conf. Multimodal Interaction Workshops, 2012,pp. 361–362.

[8] Y. Zhu, F. De la Torre, J. F. Cohn, and Y. Zhang, “Dynamic cascades with bidirectional bootstrapping for action unit detection in spontaneous facial behavior”, IEEE Trans. Affective Comput., vol. 2, no. 2, pp. 79–91, Jul. 2011.

[9] M. Dahmane and J. Meunier, “Continuous emotion recognitionusing Gabor energy filters”, in Proc. 4th Int. Conf. Affective Comput. Intell. Interaction Workshops, 2011, pp. 351–358.

[10] P. Viola and M. Jones, “Robust real-time face detection”, Int. J. Comput. Vis., vol. 57, no. 2, pp. 137–154, 2004.

[11] S. Yang and B. Bhanu, “Understanding discrete facial expressions in video using an emotion avatar image”, IEEE Trans. Syst. Man, Cybernetics Part B, vol. 42, no. 4, pp. 980–992, Aug. 2012.

[12] C.C. Chang and C.-J. Lin, “LIBSVM: A library for support vector machines”, ACM Trans. Intell. Syst. Technol., vol. 27, pp. 1–27, 2011.

[13] G. McKeown, M. Valstar, R. Cowie, and M. Pantic, “The SEMAINE corpus of emotionally colored character interactions”, in Proc. IEEE Conf. Multimedia Expo, 2010, pp. 1079–1084.