# Gesture and Scene Recognition Based Autonomous Video Lecture Acquisition Framework for Distance Learning

HAFIZ ADNAN HABIB, MUHAMMAD HAROON YOUSUF, MUID MUFTI
Telecommunication & Information Engineering Department
University of Engineering & Technology
Taxila, 47050
PAKISTAN

*Abstract:* - Video lectures are considered among reliable means for delivering class room content to distant students. Such videos are recorded manually which involves considerable time and cost of human personnel. This paper presents a novel framework for the recording of such videos. Gesture recognition and scene analysis are carried to locate the intended area of interest in the scene, to be captured by the video camera. Master-slave camera architecture is proposed for practical implementation of the system. Gesture recognition and scene analysis algorithms are applied to the video of master camera to locate intended area of interest. This information is then passed to slave camera to capture the video.

*Key-Words:* - video lecture, gesture recognition, autonomous recording, distance learning, intelligent video recording, content aware video recording.

## 1 Introduction

Distance learning is defined as a system and a process that connects learners with distributed learning resources. While distance learning takes a wide variety of forms, all distance learning is characterized by the following: separation of place and/or time between instructor and learner, among learners, and/or between learners and learning resources, and interaction between the learner and the instructor, among learners, and/or between learners and learning resources conducted through one or more media [1]. The incorporation of video and audio technologies so that students can "attend" classes and training sessions that are being presented at a remote location. Distance learning systems are usually interactive and are becoming a highly-valuable tool in the delivery of training and education to widely-dispersed students or in instances where the instructor cannot travel to the student's site [2]. Distance learning is also considered as full duplex system because students and teacher can communicate with each other.

Gesture recognition is defined as the technology transforming the movement of human body into meaningful information. Gesture recognition has been successfully applied in number of applications: sign language understanding, human computer interaction etc. Human body, partially or wholly, is analyzed and interpreted into meaningful information. This meaningful information is then utilized in some further decision making. Gesture recognition is divided in two major parts: feature extraction and recognition. Feature extraction stage extracts meaningful information from the captured gesture. These features are utilized at next stage in recognition. Statistical, neural network and temporal correlation methods are utilized at this stage. Nearest neighbor classifier, k-means classifier etc. are examples of statistical classifier. Multi-layer percepton and radial basis functions based designs are example of neural network methods. Gesture appears both static and dynamic in nature. Static gestures are referred to as postures. However dynamic gestures prolong in time. Their information is available in the video sequence. Time is also a variable in such dynamic gestures. Therefore, time factor is also considered in the representation of such gestures. This is why; dynamic gestures are also called as temporal gestures or signals having temporal nature.

## 2 Problem Formulation

Students learning takes place by the use of technology such as audio tapes, internet, web based media, interactive video conferencing and other video broadcast methods. Technological implementation of distance learning task can be divided in two major tasks: content acquisition and content delivery. Audio and video acquisition is the main components of content acquisition while

content delivery is based upon delivering and presenting the contents in an organized way. Both activities of content acquisition and content delivery are heavily explored. A lot of options for content acquisition is already designed. For content delivery, new standards are being emerged and continuously enhanced for increased performance and reliable operation. However, it can be assumed that a satisfactory level had been achieved in both content acquisition and content delivery.

Let us consider the method of content acquisition. In general, two types of hardware are involved: audio and video acquisition. Similarly in content delivery, acquired video and audio is transmitted over the transmission media.

Currently, content acquisition for both audio and video acquisition is manual or semi-automatic in functionality. Audio requires lesser control requirements than video. It has fixed space and time control. Only it has the bi-state functionality: on-state and off-state. Output of audio system is recorded and stored. Therefore, it can be concluded that audio has 1 DOF (degree of freedom). On the other hand, video requires more control due to its more DOF. Video has six DOF. Therefore video acquisition becomes an intelligent task of controlling the up-down, right-left and distance from object to camera.

The task of controlling the functionality of video acquisition is manually controlled and semi-automated depending upon the application. Traditionally, the process of video acquisition is manual. Human labor is involved in decision making for selection of intended area to be captured. A camera is mounted at some location and position and focus of camera is changed by the human operator. Semi-automatic functionality is available in some cameras. This functionality is available in terms of time selection or acquisition system changes its position at constant speed. These options are pre-programmed and usually certain combinations are available. However there is not a full fledge controlling system of acquisition system. This paper presents a framework to acquire the video the in a fully automatic function. System will initialize itself and will control the functionality of system in terms of right/left, up/down position and focus. System incorporates a video acquisition hardware, scene analysis and gesture recognition methodology to specify the intended area of scene to be captured.

# 3 Problem Solution

## 3.1 Hardware Setup

Gesture and scene recognition based framework for autonomous video lecture acquisition method has been implemented over specialized hardware. This specialized hardware consists of two cameras. In this arrangement: one camera is called master camera while the other camera is called slave camera.

Specification of master camera is as follows: fixed FOV (field of view), WFOV (wide field of view). This master camera acquires the whole scene of the lecture presentation area which includes teacher, white board and multimedia presentation area. The video of master camera is acquired and utilized in the analysis of the scene.

Specifications of slave camera are as follows: variable (FOV) camera mounted over a moveable platform in both horizontal and vertical direction. A motorized lens is installed over the camera. Lens position can be adjusted electronically. This motorized lens allows focusing / defocusing of the acquired scene. This variable field of view camera, so called slave camera, is installed over a moveable platform. This platform was designed using stepper motors connected through parallel port of the system. One stepper motor is responsible for movement in horizontal direction while the second stepper motor is responsible for controlling the motion of motorized platform in vertical direction. Therefore, slave camera has 6 DOF (degree of freedom) due to one motorized lens and horizontal and vertical stepper motors.

Video captured by the master camera is analyzed. Gesture recognition and scene analysis are preformed. Intended area of acquisition is defined in the master camera domain. This intended area is mapped from master camera coordinates into slave camera coordinates assuming the following condition.

$$S(x,y,t) \in M(x,y,t) \qquad (1)$$

Where $S(x,y,t)$ represents the video acquired by the slave camera and $M(x,y,t)$ represents the video acquired by the master camera. $x$ and $y$ represents the pixel coordinates and $t$ represents the time axis. (1) Describes that pixel information content of slave video will always be a subset of pixel information of master video. Master camera captures the video of whole presentation area which includes teacher, white board and multimedia presentation area. This acquired video is analyzed

and intended area of scene is selected based upon gesture recognition and scene analysis. Coordinates of selected area in the master video is recorded. Centroid of this area is located by following equation

$$C_{(x,y)} = median\{(x,y) \mid A\{x,y\}\} \quad (2)$$

Where $C_{(x,y)} = \begin{bmatrix} x \\ y \end{bmatrix}$ represents centroid of the set of pixels contained in the area $A(x,y)$ and

$$A\{x,y\} = \{\text{set of pixels of intended area}\}$$

and

$$A\{x,y\} \in M(x,y,t)$$

It means that $A(x,y)$ contains the instantaneous information of the master camera video $M(x,y,t)$. In addition to $C_{(x,y)}$ signal, bounding box information is computed. Bounding box [3] is defined as the smallest rectangle (oriented along any direction) that encloses the shape.

$$B = \text{bounding box}\left(A\{x,y\}\right) \quad (3)$$

Where

$$B = \begin{bmatrix} b_1 \\ b_2 \\ b_3 \\ b_4 \end{bmatrix}$$

and

$b_i = \begin{bmatrix} x \\ y \end{bmatrix}$ for $i = 1,2,3,4$ and $x, y$ describes the pixel coordinate of $b_i$. Each $b_i$ represents a point of the bounding box and there are total four points.

The recorded information of $C_{(x,y)}$ and $B$ are passed to slave camera system.

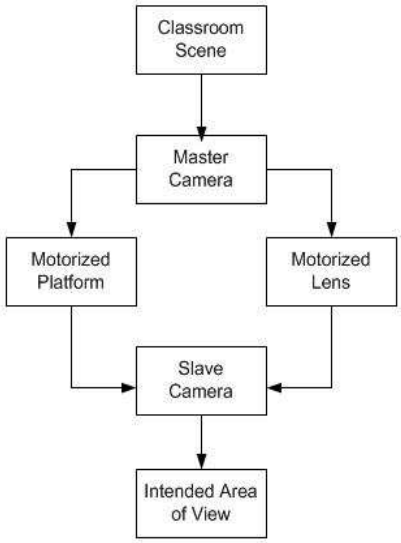
` $C(x,y)$ is utilized by stepper motors of motorized platform.

$$[O_h, O_v] = T(C(x,y))$$

Where $O_h$ and $O_v$ are offsets command given to horizontal and vertical deflection stepper motors. $O_h$ and $O_v$ are real number signals ranging from –ive maximum to +ive maximum to generate right/left or up/down motion. $(T(C(x,y))$ is the transformation which generates the offset commands $O_h, O_v$ for horizontal and vertical deflection motors. $B$ is utilized by motorized lens to adjust the filed of view in the following way

$$L = T(B)$$

Where $B$, the bounding box, is described in terms of pixel coordinates, $T$ is a transformation which converts $B$ information into the commands readable for motorized lens. $L$ is the signal which is passed to motorized lens having the range from –ive to +ive for zoom-in and zoom-out functionality.




### 3.2 System Initialization

When the system is turned on, it registers the contents of the recorded scene acquired by the master camera.

$$\{T,P,W\} = M(x,y,t) \quad (4)$$

Where

$T$ = {pixels contained by the teacher},

$P$ = {pixels contained by the multimedia presentation} and

$W$ = {pixels contained in the white board area}

$M(x,y,t)$ is the instantaneous signal by acquiring for any $t$.

### 3.3 Gesture Recognition

Video acquired by master camera is analyzed and $O_h, O_v$ and $L$ information are passed to slave camera motorized platform and motorized lens. Posture and temporal signals are developed by fixing $t$ for first signal and applying windowing function of specified length for the second signal. There are pre-recorded postures and temporal signals in the knowledge based which are used to analyze. Posture recognition is performed by Euclidean distance method [4] while temporal

information is analyzed by temporal motion templates and HMM (hidden markov model) [5][6][7][8].

## 4 Conclusion

Distance learning content acquisition is done manually or semi-automatically. This paper presents a novel concept of automating the content acquisition procedure. Here the system is proposed for class room lecture recording. However, this concept may be tested for various situations in professional, industrial, and real time image / video acquisition.

*References:*

[1]. www.trainingfinger.org/CDC_lingo.htm
[2]. www.ohsu.edu/vcs/glossary.
[3]. Castleman 1996.
[4]. Hector Hugo Aviles, Luis Enrique Sucar, Carlos Eduardo, Blance Vargas, "Visual Recognition of Gestures Using Dynamic Naïve Baysian Classifier", Proceedings of the IEEE International Workshop on Robot and Human Interface, California, USA, Oct 31- Nov 2, 2003.
[5]. Sanjay Kumarm Dinesh Kant Kumar, Arun Sharma, Neil McLachlan, "Visual hand gesture classification using temporal motion templates". Proceedings of 10th IMM'04, 2004.
[6]. Bengio, Y. "Markovian models for sequential data". Neural Computing Surveys 1999.
[7]. Ghahramani, Z. "An introduction to Hidden Markov Models and Baysian networks" International journal of Pattern Recognition and Artificial Intelligence p 9-42, 2001.
[8]. Rabiner L. R. "A tutorial on Hidden Markov Models and selected applications in speech recognition " p 4-16, IEEE Acoustics, Speech and Signal Processing Magazine 1986.