A16-199

Application Group

誰答腔我就拍誰
Intelligent Camera With Sound Direction Detection and Face Recognition

隊伍名稱　聲音魔法師 / Sound Magician
隊　　長　許哲維 / 逢甲大學電聲碩士學位學程
隊　　員　王小迪 / 逢甲大學機械與航空博士學位學程
　　　　　林谷昇 / 逢甲大學電聲碩士學位學程

指導教授

陳冠宏 / 逢甲大學電子工程學系

中正大學電機工程博士，2007 年 2 月起於逢甲大學服務。曾分別於 2013 年和 2015 年暑期至美國俄亥俄州立大學和中央研究院進行訪問研究，並於 2016 年暑期前往日本東京大學進行訪問研究。

研究領域：以視覺和聽覺為主的智慧感知運算及其加速器設計。

黃錦煌 / 逢甲大學機械與電腦輔助工程學系

美國西北大學機械工程學系博士，過去曾任美國賓州大學聲學研究所訪問學者、逢甲大學工學院院長。現任逢甲大學終身特聘教授、逢甲大學產學合作處長。

研究領域：電聲微機電系統、電聲系統設計、振動與噪音抑制。

作品摘要

近年來家用物聯網、智慧機器人、無人機與虛擬實境蓬勃發展，影像處理在這些應用上面的發展已經趨近成熟，但由於影像品質和物理上光學的限制，並不是任何情況都適用，所以我們還是需要其他各類不同的感測器來輔助。聲學感測器就是一個很好的輔助工具，聲音訊號的擷取除了可以提供發聲源在頻域上的能量資訊之外，在時域上更可以提供我們相位的資料，藉此我們可以得到聲源的方位，並與影像達到更全面的感知功能。又因半導體技術的進步，各種架構的處理器效能快速成長，使得聲音的數位訊號處理得以突破過往硬體資源上的限制，達到體積小、高運算量、多通道、精確和快速反應的優點，讓聲音的應用更加無遠弗屆。

由於全球化商業合作模式的關係，遠距離會議視訊的產品逐漸成為了不可或缺的工具。然而，多人視訊會議常常受限於開會人數、場地的影響，太多人參與或場地過大的會議，因為不方便移動視訊鏡頭的關係，可能會使一些人超出鏡頭的廣角端，甚至使大家無法聚焦在當下發言的人身上。

在本作品 Beamforming 的架構下使其對於 1k 至 5kHz 的聲源有較佳的精確度，適合於偵測人類聲音的頻響範圍。在音訊處理上使用 NI MyRIO 做為數位訊號擷取平台，此平台集成了 Xilinx Field Programmable Gate Array (FPGA) 與 ARM 處理器與多通道數位輸入輸出，方便開發者快速開發與驗證。聲音訊號經由 16 個工作時脈為 2.5MHz 之環形 Pulse Density Modulation (PDM) 麥克風陣列所組成，在 FPGA 端經由 3 階－ Cascaded Integrator–Comb(CIC) 濾波器對於 PDM 進行解碼並降低 32 倍採樣率，接著加入兩個無線脈衝 (IIR) 濾波器消除高頻雜訊與低頻所造成的直流偏移。在完成音訊的擷取後，將訊號結合麥克風位置之空間向量資訊來運算 Direction of Arrival(DOA) 得到聲源角度，接著將根據角度在 FPGA 端輸出 Pulse-Width Modulation (PWM) 來控制連續伺服馬達轉向，其中使用 Gyro Sensor 在 RT 端 (ARM) 經由 UART 協定的資料擷取，回傳當前角度來同步連續伺服馬達之轉動角度。在視訊鏡頭傳回影像後，我們同時在 PC 端使用 Color Pattern Match 的方式，輸入對照組訓練影像去做人臉辨識，並做 Moving Tracking，針對所設定的邊界閥值，做人臉置中的校正。最後，整合系統在 LabVIEW 上成為人性化的 Graphical User Interface(GUI)，加入聲控方式啟用與關閉人臉辨識或聲源偵測功能，方便使用者更直覺，更方便操作於各種情境與場合之中。本作品結合聲音與影像的追蹤與偵測，有助於提升視訊會議的品質和效率。

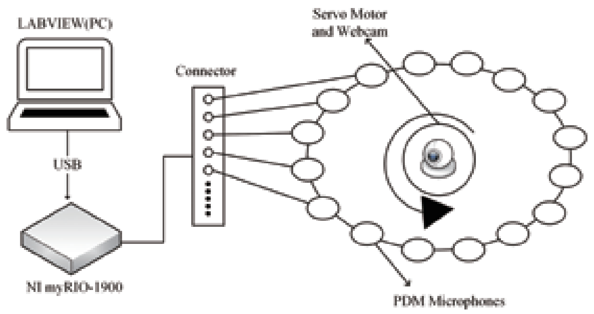

圖 1. 系統示意圖

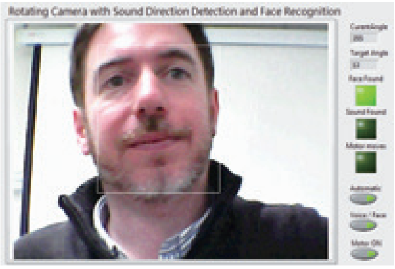

圖 2. 操作介面 (GUI)

Abstract

In recent years, there has been increased interest in home automation, intelligent robots, unmanned aerial vehicles, communication devices and augmented reality incorporating video and audio acquisition features. In certain situations, these devices need to interact with their surrounding by finding the locations of objects or audio sources. However, the acquisition of audio signals from a single microphone cannot give enough information to find the sound source. As a consequence, if the device needs to find the object, several cameras may be used. Due to advances in semiconductor technology, a variety of architecture processors has been developed in the recent years. These allow for a faster signal processing in reduced size devices, creating more reliable and powerful audio and video applications.

Due to there being the need for global communications, video conferences have become an essential tool. However, the cameras are limited by the lenses' angle of view. The camera points to a fixed position and may not be able to address all of the speakers in the meeting room. In these situations, the meeting participants may not focus on the current speaker.

This work presents a video recording device that points the camera to the speaker by employing audio source detection and face recognition algorithms. This prototype uses NI MyRIO embedded platform, which incorporates a Xilinx Field Programmable Gate Array (FPGA) and multiple input and output digital terminals. The audio signals are first acquired by a circular array of 16 digital microphones, which are Pulse Density Modulation (PDM) modulated at 2.5MHz. The beamforming provided by the position of the microphones allow for an accurate sound detection of 1k to 5k Hz. This frequency range can perfectly give an accurate detection of the human voice. The acquired signals are then real-time demodulated by a 3rd order Cascaded Integrator–Comb (CIC) and two Infiniti Impulse Response (IIR) filters. Once the 16 digital signals are decoded, the sound direction of arrival is then calculated to allow for a servo motor to aim the webcam to the speaker location. The motor is controlled by sending a Pulse-Width Modulation (PWM) signal from the FPGA outputs. The accuracy of the angle is controlled by a gyro sensor, which communicates with the ARM via UART protocol. Once the webcam is pointing to the speaker, the face recognition algorithm controls the camera rotation to keep the speaker inside the screen, even though no sound is produced. The face recognition uses pattern and moving tracking; which includes the eyes, mouth and lips. The configuration, control and monitoring of each part of the system is carried out by a Graphical User Interface (GUI) programmed in LabVIEW. In addition, this prototype can be controlled by voice, which allows for the users to activate or deactivate the motor movement, the sound detection and the face detection separately by using predefined voice commands.

This device also enhances the communication quality when a webcam is shared by several speakers in the same room, or in situations in which the speaker needs to move around the communication device.

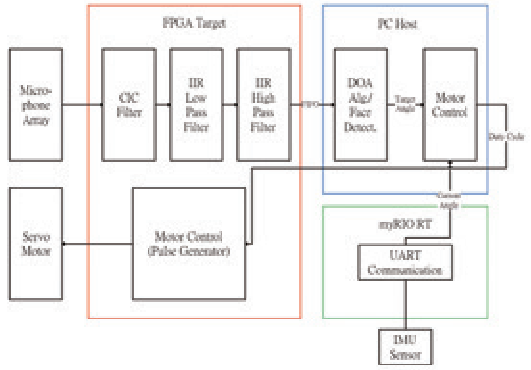

Fig 3. System structure