Multiple speaker localization and identification through multiple camera and visible light communication

Florent Lefevre, Fabián Seguel, Vincent Bombardier, Nicolas Krommenacker, Patrick Charpentier, Bertrand Petat

Multiple Speaker Localization and Identification through Multiple Camera and Visible Light Communication

Lefevre Florent *†, Seguel Fabián * ‡, Bombardier Vincent *,
Krommenacker Nicolas *, Charpentier Patrick * and Petat Bertrand †
* Université de Lorraine, CRAN CNRS UMR 7039,
F-54506 Vandoeuvre-les-nancy
(florent.lefevre, nicolas.krommenacker, vincent.bombardier, patick.charpentier)@univ-lorraine.fr
†CitizenCam, 132 rue andré Bisiaux
54320 Maxéville
(flefevre, bpetat)@citizencam.eu
‡*Universidad de Santiago de Chile*
Departamento de Ingeniería eléctrica, Avenida Ecuador N° 3519,
Estación Central, Santiago
fabian.seguelg@usach.cl

***Abstract*—We propose a novel method for the localization and the identification of a speaker in the context of conference situations. For doing this, LED lights attached to the microphones are used as visible light communications transmitters. Once the speaker activate the microphone, LED transmits the speaker identification. At the receiver side, LED lights are identified and separated from video streams acquired by a low cost IP CMOS camera. Then, identification of the speaker is performed by recovering the data from the separated image. Results in terms of BER acquired from experimental demonstration are presented for different distances between the camera and the light source, i.e., 35 cm, 1, 2 and 5 meters.**

I. INTRODUCTION

This paper presents a new approach of speaker detection and identification by analyzing video streams from a multi-camera system. The proposed method can be applied for automatic detection and identification on council meetings, international forums, public debates or any scenario where multiple potential speakers are visible on multiple video streams. Commonly, to perform speaker detection, audio signals are used to confirm the presence of a person in the image. As an example, in [1], the authors propose a speaker detection algorithm using the correlation between sound and video with a high accuracy. Nevertheless, the proposed method is useful for single speaker identification. On the other hand, in [2] a multi-microphone system allows an accurate localization of the dominant speaker when is correlated with gesturing motion in video. Nevertheless the deployment of the multiple microphones infrastructure is expensive and time demanding. In our context, multiple CMOS cameras are recording multiple speakers at the same time, but only one audio channel is available. Due to this, localization of the source through audio signal cannot be implemented. In order to detect the speakers, we offer to detect the light emitted by the microphone when a person speaks. Indeed, most of the French municipality is equipped with microphone systems where a LED lights up on the microphone when someone speaks. The detection of this light allows the detection of the speaker without changing the actual system. Furthermore, the LED light attached to microphone may be used as visible light communications transmitters.

Optical wireless communications (OWC) uses the infrared and visible spectrum of electromagnetic waves to provide wireless communication. In particular, VLC, which uses the visible spectrum has gained attention the last decade since its potential to provide wireless communications using the existing light deployment and the tremendous advance in the research of Lighting Emitting Diodes (LED). Usually, to provide VLC, light intensity is modulated with information signal. The modulated light is detected at the receiver side by photodiodes which can transform the light intensity into a proportional electrical current. Since many devices such as cellphones have incorporated cameras and flash lights, many researchers have recently focused their studies in providing communication by using these off-the-shelf components to generate low cost transceivers. Many applications of optical camera communications (OCC) have been developed for vehicular technology such as vehicle to vehicle (V2V) and infrastructure to vehicle (I2V) and positioning applications [3]. Most of the existing OCC systems have been deployed for fixed lighting and receivers [4], [5], [6]. Recently, a novel approach has been proposed in order to overcome the mobility of the receiver and/or transmitter which has been named as region of interest [7]. Since lighting is not always placed in exactly the same position within the image, a preprocessing stage is needed to determine the position of the light inside the entire frame. Once the position of the light is determined, it is separated from the whole image and the VLC data is retrieved from it.

The section II presents the overview of the proposed method whilst the section III details the algorithm for the OCC system and the algorithm used for LED localization and separation. In section IV the results of the experiment are presented and finally, in section V the main conclusions of the study and future work is delivered.

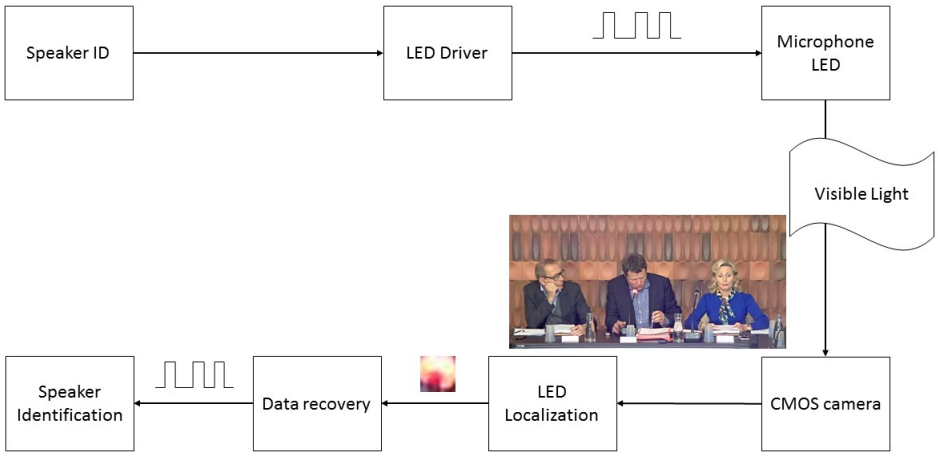

Fig. 1. Overview of the proposed method

II. System description

This paper presents a method based on optical camera communications to detect and identify speakers which can be used in conference situation using cheap CMOS cameras for transmitting on-line video streaming. For doing this, we propose the usage of an OCC system in which each speaker will have their own microphone. Microphones will be equipped with a small LED light. The LED light will transmit an unique pre-defined code for each speaker. When the microphone is active (speaker talking), LED light is turned on and it sends speakers identification code by means of UFSOOK modulation. At the receiver side, LED lights are first localized and separated from video streams acquired by a CMOS camera. Once the LED light is localized, the region of interest is used to retrieve identification of the speaker. By doing this, LED source is separated pixel by pixel from the rest of the image and identification code is obtained. When a single camera is considered, data recovery is directly performed from the separated light image. Fig. 1 shows the system diagram for the proposed method.

III. Methods

In this section the methods used in the transmitter and receiver side are presented.

A. Under sampled frequency shift for optical camera communications

Under sampled frequency shift OOK (UFSOOK) modulation [8] is used to transmit data when low sampling rate cameras are used at the receiver side. In the market, most of the low cost cameras can work with an speed of 30 frames per second. This modulation employs two special designed square wave patterns with different frequencies to represent mark (bit 1) and space (bit 0). In order to be able to recognize each different bit at the receiver side two square waves with different frequencies are used , i.e., f_{s1} and f_{s2}. Formally, it can be expressed as

$$s(t) = \begin{cases} [cos(2\pi(m+0.5)f_{camera}t)] & if\ bit = 1 \\ [cos(2\pi m f_{camera}t)] & if\ bit = 0 \end{cases} \quad with\, 0 < t < T_c \tag{1}$$

Where $[\,]$ is the square wave function of frequency $f_{S1} = (m+0.5)f_{camera}$ and $f_{S2} = m.f_{camera}$ with $m \in Z$. Since $f_{S1} = (m+0.5)f_{camera}$, the LED will have two different states in two successive images (light OFF and light ON or vice versa). On the other hand, when a zero is send, the LED oscillation will be synchronized with the camera frame rate and the same image will be seen in two consecutive camera shots.

B. LED Localization

Since in some conference situation the position and the state of the light source may change, we have to frequently check the state of the LED light in the image (ON or OFF) and where its position is. The method presented in this paper detects the light attached to a microphone for doing a segmentation of the entire frame in which only the LED light will be considered. After this, data retrieval is performed by used the segmented image. The principle of our method is shown in Fig. 2 where each process is executed sequentially. In order to use this method in every situation, we introduce an initialization step to manually select the limits of the research region and the HSV threshold.

Selection of the Region of Interest (ROI): The first step of our method is the selection of the ROI involving a reduction of computation's time and the risk of wrong detections.
Since microphones are located on the table in front of every

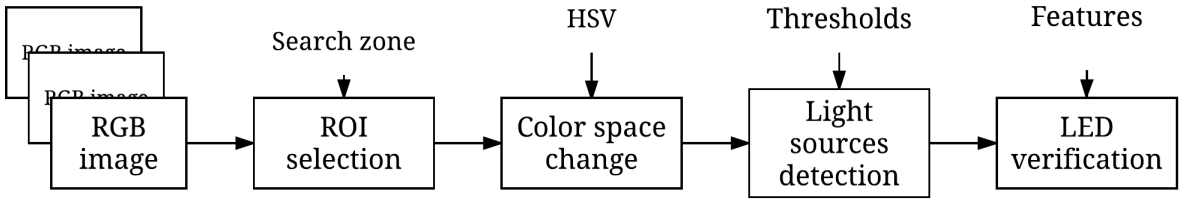

Fig. 2. LED localization method

eventual speaker, it is appropriate to look for microphones between the table and the top of the head of speakers. Because the microphones are not fixed on the tables, we need to define a large area in case the speakers move the microphones.

Color space selection: The second step of the method is to prepare for the light detection stage. To do this, it is necessary to change the color representation. Different color spaces were tested (RGB, HSV [9], CIE L*a*b* [10]). Those color spaces, frequently used in colorimetry, allow an efficient representation of the luminosity in the image, especially with the Intensity component of the HSV color space or the Luminance component of the L*a*b* color space.The choose of the HSV color space was made in line with the features we've selected for the LED verification step.

Detection of light sources: In the first place, we want to localize microphones which are currently used. In other words, we want to find light sources emitted by active microphones. The HSV color model and especially the V component allows to efficiently find them. We execute thresholding to get candidate regions, as we can see in the following equation :

$$C(x,y) = \begin{cases} 1 & if \quad ((S(x,y) \geq Ts1 \cap S(x,y) \leq Ts2) \\ & \cap (V(x,y) \geq Tv1) \cap V(x,y) \leq Tv2)) \\ 0 & else \end{cases} \quad (2)$$

Where $C(x,y)$ is the result of the thresholding operation, $S(x,y)$ and $V(x,y)$ are the saturation and intensity values. The four threshold values (Ts1,Ts2,Tv1 et Tv2) are empirically defined.

The lighting condition may change during the videos, therefore many reflections may occur, resulting in small lighting sources in the thresholded image. We apply a connected component analysis [11] in order to execute a dimensional thresholding. The light sources which are inferior (like reflections) or superior (like lighting) to the light source from a microphone are ignored.

Speaker verification: In order to check that the light sources are from microphones' LED, we use a classification tree [12] and a sliding windows techniques to separate zones that contain one active microphone from another. For each light sources, a windows of size 19x19 , is swept across the candidate regions. The features calculated in each window location are then tested with a classifier. The classification tree is created during the initialization step of our system. Using a classification tree was a choice made thanks to the possibility of interpreting the causal connection, unlike methods like neural network [13], KNN [14] or SVM [15]. We use seven features (mean, variance and third central moment of the H component, mean of the S component, as well as the mean, variance and the V component's Kurtosis fourth root) that make it possible to obtain a microphone model that can be used for different cameras but have close viewing characteristics.

IV. Results

To test the performance of the prototype, we use an AXIS Q-3505 CMOS camera. The frame rate of the camera is set to 30 fps and the resolution used in the experiments is 1280x720. The system is evaluated by transmitting a short message, i.e. the identification of the speakers, with different distance between the LED and the camera. At the receiver side, a small 3W LED light is used in order to transmit the VLC code. LED light is connected to an Arbitrary Waveform Generator (AWG) AGILENT 33120A as shown in Fig. 3.

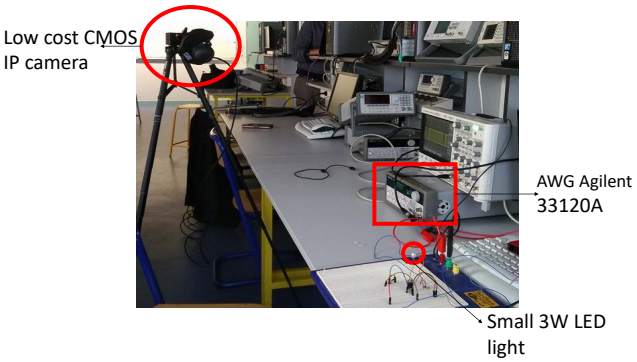

Fig. 3. Experimental set up

The experiments where performed using three different distances, i.e., 35 centimeters, 1 , 2 and 5 meters. The microphone code sent though UFSOOK was MIC1 and MIC2 in order to differentiate two different microphones. Since the frame rate of the CMOS IP camera was set to 30 fps, two different frequencies where used to signaling bit 1 or bit 0. f_{s1} and f_{s2} were set to 105 and 120 Hz respectively for bit=1 and bit=0.

A Frame header of frequency 10 kHz was used before the payload. The data sent by the AWG is shown in Fig. 4

Fig. 4. Transmitted data from AWG

In Table I the bit error rate (BER) obtained for each experiment are shown. The message was sent and recorded multiple times.

TABLE I
IMPACT OF THE DISTANCE (IN M) ON THE BIT ERROR RATE (10^{-3})

Distance (m)	0.35	1	2	5
BER (10^{-3})	8.7	22.2	9.52	9.92

V. Conclusions

In this paper a method for speaker identification and localization based on optical camera communications was presented. The proposed method obtains the region of interest by means of a sequential process. After the region of interest is separated from the frame, the demodulation of the UFSOOK signal is performed by matching two consecutive images. The proposed system show have a good performance even for long distances, i.e., 5 meters. Due to this, it is possible to provide a real on-line system for video transmission which uses OCC in order to determine the identification and the position of the speaker automatically. For future work, different modulation schemes will be tested for low sampling rate cameras as well as different LED light configurations that can be attached to microphones.

Acknowledgment

This work was funded by the Beca Doctorado Nacional 2016 CONICYT PFCHA/21161397

References

[1] R. Cutler and L. Davis, "Look who's talking: speaker detection using video and audio correlation," in *2000 IEEE International Conference on Multimedia and Expo. ICME2000. Proceedings. Latest Advances in the Fast Changing World of Multimedia (Cat. No.00TH8532)*, vol. 3. IEEE, pp. 1589–1592.

[2] E. D'Arca, N. M. Robertson, and J. R. Hopgood, "Look who's talking: Detecting the dominant speaker in a cluttered scenario," in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, may 2014, pp. 1532–1536.

[3] N. T. Le, M. S. Ifthekhar, Y. M. Jang, and N. Saha, "Survey on optical camera communications: challenges and opportunities," *IET Optoelectronics*, vol. 9, no. 5, pp. 172–183, 2015.

[4] C. Danakis, M. Afgani, G. Povey, I. Underwood, and H. Haas, "Using a CMOS camera sensor for visible light communication," in *2012 IEEE Globecom Workshops*. IEEE, dec 2012, pp. 1244–1248.

[5] W. Huang, P. Tian, and Z. Xu, "Design and implementation of a real-time CIM-MIMO optical camera communication system," *Optics Express*, vol. 24, no. 21, p. 24567, 2016.

[6] P. Luo, Z. Ghassemlooy, H. Le Minh, X. Tang, and H.-M. Tsai, "Undersampled phase shift ON-OFF keying for camera communication," in *2014 Sixth International Conference on Wireless Communications and Signal Processing (WCSP)*. IEEE, oct 2014, pp. 1–6.

[7] T. Nguyen, A. Islam, and Y. M. Jang, "Region-of-Interest Signaling Vehicular System Using Optical Camera Communications," *IEEE Photonics Journal*, vol. 9, no. 1, 2017.

[8] R. D. Roberts, "Undersampled frequency shift ON-OFF keying (UFSOOK) for camera communications (CamCom)," in *2013 22nd Wireless and Optical Communication Conference*. IEEE, may 2013, pp. 645–648.

[9] G. H. Joblove and D. Greenberg, "Color Spaces for Computer Graphics," in *Proceedings of the 5th Annual Conference on Computer Graphics and Interactive Techniques*, ser. SIGGRAPH '78. New York, NY, USA: ACM, 1978, pp. 20–25.

[10] I. 11664-4, "ISO 11664-4: 1976 L* a* b* Colour Space," *Joint ISO/CIE Standard, ISO*, pp. 11 664–4, 2008.

[11] C. Fiorio and J. Gustedt, "Two linear time Union-Find strategies for image processing," *Theoretical Computer Science*, vol. 154, no. 2, pp. 165–181, Feb. 1996.

[12] J. R. Quinlan, "Induction of decision trees," *Machine Learning*, vol. 1, no. 1, pp. 81–106, Mar. 1986.

[13] H. A. Rowley, S. Baluja, and T. Kanade, "Neural network-based face detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, no. 1, pp. 23–38, Jan. 1998.

[14] H. Zhang, A. C. Berg, M. Maire, and J. Malik, "SVM-KNN: Discriminative Nearest Neighbor Classification for Visual Category Recognition," in *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, vol. 2, 2006, pp. 2126–2136.

[15] E. Osuna, R. Freund, and F. Girosit, "Training support vector machines: An application to face detection," in *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Jun. 1997, pp. 130–136.