Unconventional TV Detection using Mobile Devices

Mohamed Ibrahim, Ahmed Saeed
and Moustafa Youssef
Department of Computer Science and Engineering
Egypt-Japan University of Science and Technology(E-JUST)
{mibrahim,ahmed.saeed,moustafa.youssef}@ejust.edu.eg

Khaled A. Harras
Computer Science Department
School of Computer Science
Carnegie Mellon University Qatar
Email: kharras@cs.cmu.edu

***Abstract*—Recent studies show that the TV viewing experience is changing giving the rise of trends like *"multi-screen viewing"* and *"connected viewers"*. These trends describe TV viewers that use mobile devices (e.g., tablets and smart phones) while watching TV. In this paper, we exploit the context information available from the ubiquitous mobile devices to detect the presence of TVs and track the media being viewed. Our approach leverages the array of sensors available in modern mobile devices, e.g., cameras and microphones, to detect the location of TV sets, their state (ON or OFF), and the channels they are currently tuned to. We present the feasibility of the proposed sensing technique using our implementation on Android phones with different realistic scenarios. Our results show that in a controlled environment a detection accuracy of 0.978 F-measure could be achieved.**

***Index Terms*—TV detection; ubiquitous sensing; mobile computing; audio fingerprinting; computer vision**

I. Introduction

TV viewers' profiling is an important functionality for both advertisers and service providers. Traditionally, the detection techniques of TV viewers' habits are concerned more about the collective preferences of the viewers and rely mainly on focus groups [16] or special hardware connected to the TV (e.g., set top devices) [19]. Recent studies show that 52% of cell phone owners use their phones while watching TV [17] and 63% of tablets owners use their tablets while watching TV [1] in what was called *"Connected Viewers"*. The rise of these *"Connected Viewers"* opens the door for a new unconventional approach for TV viewers' profiling based on the ubiquitous mobile devices and their equipped sensors. Such approach can provide ubiquitous fine-grained information about the user's TV viewing preferences leading to new possibilities for advertisers and service providers on both the TV and mobile sides.

In this paper, we present the design, implementation and evaluation of a system that can leverage the large array of sensors currently available in smart phones and other mobile devices to accurately detect the presence of TV sets. In particular, our implementation currently depends on the acoustic context analysis and visual surroundings detection using microphones and cameras embedded in mobile devices to identify (1) the presence and locations of the TV sets, (2) whether they are ON or OFF, and (3) the channels they are currently tuned to. We test the system's ability to differentiate between a TV set's acoustic and visual fingerprints on one side and other sources of similar fingerprints such as people having a conversation and laptops playing audio/video files on another side. The results showed that a typical mobile device can reach an F-measure of 0.978 in a *controlled environment*.

Our goal is to develop a novel system that could be deployed on mobile devices to passively detect the TV viewing preferences of the mobile devices owners. This TV viewing history information, analogous to web browsing history information, can assist with enhancing the user mobile and web browsing experience. This information can also give way to new social applications that connect users with the same TV viewing preferences. On the other hand, this information will be invaluable to advertisers by providing fine-grained audience measurement, tracking mobile users' preferences through their TV viewing habits which can enable a new generation of targeted mobile ads and more informed planning of TV ads.

This approach sets itself apart from earlier work that detects TV shows or identifies playing music based on acoustic fingerprints [2], [7], [8], [14], [20] by allowing for passive detection of TVs and the shows they are playing. Conventional popular applications (e.g. IntoNow [2]) are interactive applications that require the user to operate it to detect the TV show playing which assumes the presence of an operating TV set. This approach is not appropriate for a passive application that aims at tracking the user's TV view preferences. Moreover, all these audio detection approaches focus on identifying the content regardless of its source and hence cannot determine the audio source type (whether it is a laptop, people talking, or TV). On the other hand, our proposed system addresses the challenges of detecting the presence of a TV set and determining whether it's turned on or off before determining the show playing.

The rest of the paper is organized as follows: In the next section, we provide our vision and architecture for the novel sensing approach. Section III presents and evaluates our mobile TV detection service. Then, we briefly discuss related work in Section IV. Finally, Section V discusses the challenges and gives directions for future work.

II. System Architecture

Figure 1 shows the proposed architecture. Standard mobile devices with embedded sensors (such as mic, camera, accelerometer, Global Positioning System (GPS), etc.) submit their location tagged sensory information to the system's server. The server has three main components: Mobile TV Detection Service, TV Detection Supporting Services and TV Detection Applications.

 ISBN: 978-1-61208-289-9

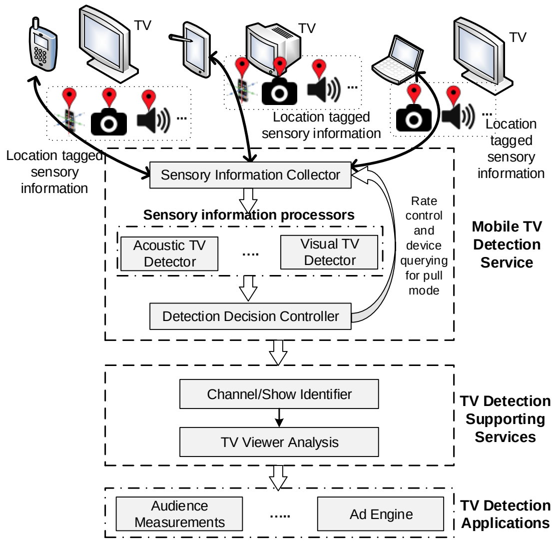


Fig. 1. System architecture.

Mobile TV Detection Service is responsible for the collection and processing of the sensory information. This service contains different modules responsible for the detection of TV sets based on information collected from different sensors. It is also responsible for the fusion of the detection decision made by the different sensors. Moreover, this service is responsible for controlling the rate at which the sensors collect their information.

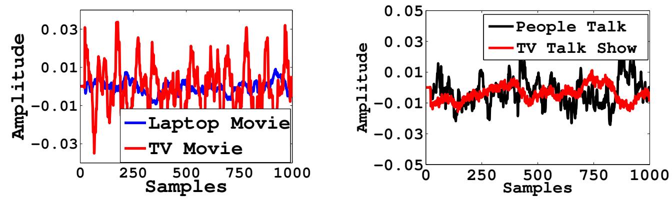
TV Detection Supporting Service is responsible for further processing of the information collected about the detected TV sets. It connects to TV streaming servers, online schedules, and media databases to detect the current channel. It depends on the comprehensive previous techniques for detecting TV shows, music, commercials, channels [3], [5], [6], [14], [20]. Other possibilities include interaction with social network sites to access information about the user preferences.

TV Detection Applications use the TV sets information collected by other services to provide different services either to the mobile user (e.g., personalization) or to third party applications (e.g., audience measurement systems and targeted ads systems).

For the rest of the paper, we focus on the detection of the presence of TV sets using mobile phones. We present the design, implementation and evaluation of the *Mobile TV Detection Service*.

III. Mobile TV Detection Service

We implemented the service on different Android phones and used it while watching different TV sets made by different manufacturers. We tested our implementation in a *controlled environment* using two sensors: microphone and camera. We address the challenge of differentiating the visual and acoustic signature of TV sets and other sources that could be confused with the TV. For example, the sounds coming from a TV set could be confused with a laptop or normal people talking.



(a) Movie on a TV and laptop

(b) TV show and normal people talking

Fig. 2. Acoustic time-domain raw data amplitude.

Moreover, the visual signature of a TV set (i.e., rectangular-shaped object with varying content) could be confused with picture frames and windows.

A. Acoustic TV Detector

The main challenge for acoustic TV detection is extracting unique features for the acoustic fingerprint that would enable the differentiation between TV sets and other sources that could be confused with it. We collected an acoustic dataset composed of 91 audio recordings for training and 60 independent audio recordings for testing. Each audio recording is 30 seconds long. We had different configurations including the TV volume, phone relative distance to the TV, position of the phone (in pocket, on couch, etc.), show type (movie, talk show, sports, etc.), gender and talking level of the actor/anchor. Also, we collected a data set under the same different configurations for the laptop and normal people talk classes. Our goal in the rest of this section is to identify time and frequency domain features that can differentiate between the TV case on one hand and the laptop and people talking case on the other hand. Figures 2 and 3 show sample raw data obtained from our acoustic dataset.

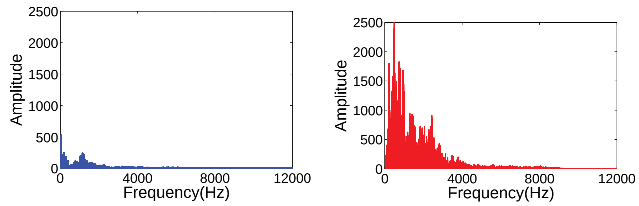
1) Time domain features: Figure 2(a) shows the raw time domain acoustic amplitude for listening to a movie on a TV and on a laptop, whereas Figure 2(b) shows the same signal while listening to a TV show and listening to a group of people talk. The figure shows that there is a noticeable difference between the two cases in each figure. This is intuitive, as a person listening to a movie or show on a laptop will usually have a lower volume than the case of listening to the same show on the TV. On the other hand, people talking will tend to lower the volume of the TV.

Based on Figure 2, we extract features that capture the amplitude variations of the acoustic signal in the time domain. One of these key features is the Zero Crossing Rate (ZCR) [13] that represents the rate at which the signal crosses the x-axis (zero amplitude).

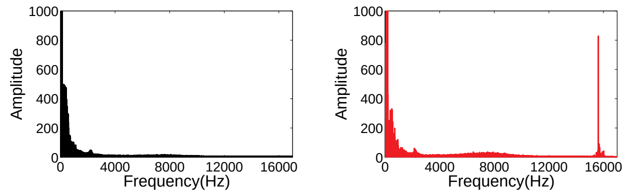
$$ZCR = \frac{1}{N}\sum_{i=1}^{n} |sign(data(i)) - sign(data(i-1))|$$

$$\begin{aligned} sign(x) &= 1, x > 0 \\ sign(x) &= 0, x = 0 \\ sign(x) &= -1, x < 0 \end{aligned} \quad (1)$$

ISBN: 978-1-61208-289-9



(a) Watching a movie on laptop vs. TV



(b) Normal people talk vs. Watching a talk show on TV

Fig. 3. Acoustic frequency-domain raw data.

ZCR is used to estimate the fundamental frequency of the signal. Therefore, it is used as an indication of the noisiness of the signal. Another time domain feature is the Short Time Energy (STE) [13] that captures the loudness of the signal and is computed as the average of the square amplitude of the signal:

$$STE = \frac{1}{N}\sum_{i=1}^{n} data^2(i) \tag{2}$$

2) Frequency domain features: Figure 3 shows the frequency domain signal for the same example as in Figure 2. The figure shows that the frequency domain response of the signal differs from the TV and other classes. From the figure, it could be observed that media streamed to laptops are lower quality in terms of bit rate compared to media displayed on the TV. This observation leads to the conclusion that the acoustic fingerprint of laptops will have a lower bandwidth as compared to TV sets. Similarly, comparing the acoustic fingerprint of a TV set and normal people talk, it could be observed that the TV set's fingerprint is a combination of people talk (4 KHz) and other sources (e.g, music (16 KHz)). This observation also leads to the conclusion that people conversations will have a lower bandwidth as compared to TV sets in the frequency domain. Based on these observation, we use the following frequency domain features: Spectral Centroid (SC) and Spectrum Spread (BW) [13].

$$SC = \frac{\sum_{i=1}^{n} f(i) * A(i)}{\sum_{i=1}^{n} A(i)} \tag{3}$$

$$BW = \frac{\sum_{i=1}^{n} (f(i) - SC)^2 * A(i)}{\sum_{i=1}^{n} A(i)} \tag{4}$$

where $f(i)$ is frequency and $A(i)$ is the amplitude values at index i. These features represent the spectrum by its center of mass and its spread around that center. We also use the Mel-frequency Cepstral Coefficients (MFCC) [13] which are a set of features, where each feature represents a portion of the spectrum in the Mel scale.

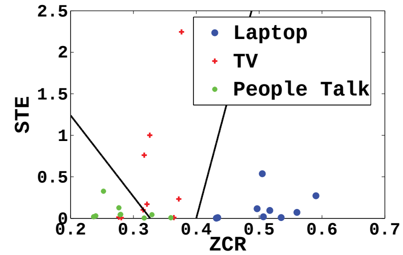


Fig. 4. SVM discriminant function using two features.

3) Acoustic fingerprint classification: After extracting the features, we use a Support Vector Machine (SVM) classifier to distinguish TVs from the other two classes. Figure 4 shows a sample result using the classifier for two features (ZCR and STE). As the figure shows, the three classes are linearly separable, except for some TV talk shows and people talk samples, then the SVM classifier, which is a discriminate based classifier, can easily classify them.

B. Visual TV Detector

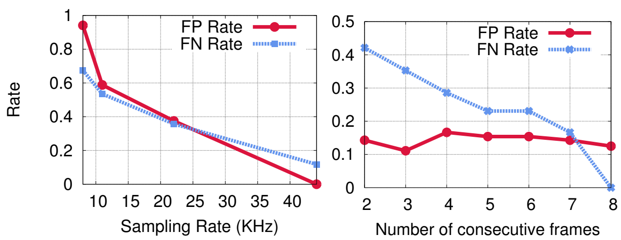
Acoustic detection may confuse the presence of TV sets with other sources of similar acoustic fingerprints, e.g., stereo players. To reduce this ambiguity, we consider the usage of cameras as a source of detection information. Our approach is based on recent statistics that show that a smart phone user holds the smart phone at 12.5 inches from her eyes while surfing and at 14.1 inches away from her eyes while texting [4]. At these distances, if the smart phone user is watching the TV, the TV will either partially or fully appear within the camera's frame. We collected 26 shots by normal users using their phones, e.g., to text or browse the Internet, with each shot composed of 8 consecutive frames. Fourteen out of the twenty six shots were taken in different light conditions in different locations with a mix of shots showing the TV as a whole or partially. The remaining 12 shots had no TV sets but rather objects that could be confused with TV sets using our proposed algorithm (e.g., windows, doors, picture frames, and shots with moving people and objects).

We use a simple algorithm that detects the characteristics of a TV in a sequence of frames captured either through a recorded video or sequence of captured images. The algorithm works in three steps, summarized in Figure 5. The first step, Figure 5(a), detects changing parts in the image sequence, which represent the dynamics of the scene on a TV set. This is performed using a simple background extraction algorithm [22]. In the second step, Figure 5(b), it determines rectangle-shaped contours within each image filtering out small contours (smaller than 5% of the total area of the image) and large contours (larger than 70% of the total area of the image). The rectangle shapes detection works in two steps: the first finds all contours in the image using the algorithm proposed in [18]. In the second step, all contours are simplified by reducing

Copyright (c) IARIA, 2013. ISBN: 978-1-61208-289-9

(a) The changing areas between different frames are detected as foreground.
(b) All rectangles in picture are detected then small rectangles and large rectangles are filtered out.
(c) The intersection of the previous two steps is performed to detect the existence of the TV.

Fig. 5. TV detection steps using the camera.



(a) Audio sampling rate effect on detection accuracy.
(b) Camera frames number effect on false positive and false negative rates.

Fig. 6. Summary of Individual Sensor Results

TABLE I
COMPARISON BETWEEN DIFFERENT TV DETECTION APPROACHES.

Approaches	Acoustic	Visual	**Fused**
False Negative Rate	0.13	0	**0**
False Positive Rate	0	0.125	**0.042**
F-measure	0.928	0.933	**0.978**

the number of points forming them using the Ramer-Douglas-Peucker algorithm [9]. Reduced convex contours composed of only four points are selected as rectangle-shaped areas. Finally, an intersection step between the previous two stages is performed to produce the final output (Figure 5(c)). In particular, the rectangle with the smallest area that encloses all the recorded foreground contour centers is declared to be a TV.

C. Detection Decision Controller

This module is responsible for fusing the decisions made based on different sensors. Furthermore, it is also responsible for setting the frequency by which the sensory information are collected (e.g., acoustic sampling rate and number of captured frames per second) to control both the accuracy and energy efficiency of the system. The sensors fusion is based on the assumption that avoiding false negatives is more important than false positives, as not detecting a TV set wastes opportunities of detecting context information. Therefore, we fuse the results of the audio and video modules using the a simple OR rule: If the two techniques result in two opposite results, then the fused results will always be positive, i.e., a TV is detected.

Figure 6(a) shows the effect of the acoustic sampling rate on the false positive and false negative rates. The figure shows as the sampling rate increases, more information is extracted from the audio source and lower false positive and false negative rates are achieved. Figure 6(b) shows the effect of increasing the number of consecutive frames on the visual detection algorithm. Table I summarizes the results. The acoustic approach achieves a zero false positive rate and a 0.13 false negative rate (0.928 F-measure) with most of the errors in mixing a quiet talk show on the TV with normal people talking. On the other hand, the visual detection approach achieves a detection accuracy of zero false negative rate and a 0.125 false positive rate (0.933 F-measure). The results of the fusion approach is summarized in Table I. This approach results in a zero false negative rate and 0.042 false positive rate (0.978 F-measure). Note that this can also be further enhanced by combining the detection results from different nearby devices and other sensors.

IV. RELATED WORK

Extensive work has been done in detecting real time TV shows, commercials, and channels [3], [6]. This involves scene boundary detection [15] and TV shows recognition [10]. Another line of work depends on audio as their data source for TV shows, music and channel identification [5]–[8], [14], [20]. E. Bisio et al. [6] showed how to detect in real time what people are watching on TV using audio signals recorded by their smart phones. The IRCAM audio Fingerprint framework [14] enhances the accuracy of two commercial systems: Philips [11] and Shazam [20]. In [8], a Filtering approach was proposed to extract music from background noise and identify it. However, all these audio detection approaches focus on identifying the contents regardless of its source and hence cannot determine the audio source type (whether it is a laptop, people talking, or TV).

On another perspective, earlier work investigating the detection of TV sets, e.g., [21], relied on special devices for sensing. This work detects the power leakage of a receiver's local oscillator in order to detect the presence of a TV set. This technique required the usage of special hardware that needed to be setup in the vicinity of the TV set. Such systems do not scale and are harder to deploy.

V. Conclusion and Future Work

A. User Privacy

Protecting the user privacy can be achieved by local processing of the raw data on the user mobile device and forwarding only the anonymized detection results. This can be extended by forwarding the data from the mobile phone to a more powerful device, such as the user laptop for processing before forwarding to the back end server. Another possibility is to provide an option for users to avoid visual recordings and use only acoustic recordings for better privacy settings. Moreover, secure computations can be used on encrypted data [12], which is still an area with a space for huge improvement. Attacks on the system, e.g., by submitting incorrect data, should also be accounted for.

B. Incentives

To encourage the users to deploy the proposed system, different incentive techniques can be used including providing coupons, recommendation services, among other traditional incentive systems.

C. Using Other Sensors

The proposed approach can be extended to use other sensors. For example, the inertial sensors (e.g., accelerometer, gyroscope and compass) can be used to better trigger the acoustic and visual detection sensors based on the detected user activity. Other sensors, such as WiFi, can be used to obtain the device location indoors and hence provide better context information about the user actions and the TV location.

D. Energy Efficiency

Continuous sensing on a mobile device can quickly drain the scarce battery resource. Automatically setting the sensing rate and which devices to sense based on their remaining battery, the device context and location, and required information are different steps to address this issue. This is one of the main functionalities of the *Detection Decision Controller*. In addition, offloading the computations to a more powerful user device can also help alleviate this concern.

E. Large Scale Deployments

The results presented in the paper were conducted as a proof-of-concept in a controlled environment. A large scale deployment of the application should be conducted to measure the performance of the proposed system in a real environment. A major challenge of such a large scale evaluation include determining the ground truth of whether the user is actually viewing the TV or not.

References

[1] "Tablets and Multi-Tasking," *The GfK MRI iPANEL Reporter*, 2012.

[2] "IntoNow," http://www.intonow.com/ci, Last Accessed July 2013.

[3] A. Albiol, M. Ch, F. Albiol, and L. Torres, "Detection of TV commercials," in *Acoustics, Speech, and Signal Processing, 2004. Proceedings. (ICASSP '04). IEEE International Conference on*, vol. 3. IEEE, May 2004, pp. iii–541.

[4] Y. Bababekova, M. Rosenfield, J. Hue, and R. Huang, "Font size and viewing distance of handheld smart phones," *Optometry & Vision Science*, vol. 88, no. 7, pp. 795–797, 2011.

[5] I. Bisio, A. Delfino, G. Luzzati, F. Lavagetto, M. Marchese, C. Fra, and M. Valla, "Opportunistic estimation of television audience through smartphones," in *Performance Evaluation of Computer and Telecommunication Systems (SPECTS), 2012 International Symposium on*, 2012, pp. 1–5.

[6] I. Bisio, A. Delfino, F. Lavagetto, and M. Marchese, "A Television Channel Real-Time Detector using Smartphones," *IEEE Transactions on Mobile Computing*, vol. 99, no. PrePrints, p. 1, 2013.

[7] M. Casey, R. Veltkamp, M. Goto, M. Leman, C. Rhodes, and M. Slaney, "Content-based music information retrieval: Current directions and future challenges," *Proceedings of the IEEE*, vol. 96, no. 4, pp. 668 –696, April 2008.

[8] C.-Y. Chiu, D. Bountouridis, J.-C. Wang, and H.-M. Wang, "Background music identification through content filtering and min-hash matching," in *ICASSP 2010*, 2010, pp. 2414 –2417.

[9] D. Douglas and T. Peucker, "Algorithms for the reduction of the number of points required to represent a digitized line or its caricature," *Cartographica*, vol. 10, no. 2, pp. 112–122, 1973.

[10] M. Fink, M. Covell, and S. Baluja, "Social-and interactive-television applications based on real-time ambient-audio identification," in *Proceedings of EuroITV*. Citeseer, 2006, pp. 138–146.

[11] J. Haitsma, "A highly robust audio fingerprinting system," in *ISMIR 2002*, 2002, pp. 107–115.

[12] M. Naehrig, K. Lauter, and V. Vaikuntanathan, "Can homomorphic encryption be practical?" in *Proceedings of the 3rd ACM workshop on Cloud computing security workshop*, ser. CCSW '11, 2011, pp. 113–124.

[13] G. Peeters, "A large set of audio features for sound description (similarity and classification) in the CUIDADO project," Tech. Rep., 2004.

[14] M. Ramona and G. Peeters, "Audio identification based on spectral modeling of bark-bands energy and synchronization through onset detection," in *ICASSP 2011*, May 2011, pp. 477 –480.

[15] Z. Rasheed and M. Shah, "Scene detection in Hollywood movies and TV shows," in *Computer Vision and Pattern Recognition, 2003. Proceedings. 2003 IEEE Computer Society Conference on*, vol. 2, 2003, pp. II – 343–8 vol.2.

[16] C. A. Russell and C. P. Puto, "Rethinking television audience measures: an exploration into the construct of audience connectedness," *Marketing Letters*, vol. 10, no. 4, pp. 393–407, 1999.

[17] A. Smith and J. L. Boyles, "The rise of the Connected Viewer," *Pew Internet & American Life Project*, 2012.

[18] S. Suzuki *et al.*, "Topological structural analysis of digitized binary images by border following," *Computer Vision, Graphics, and Image Processing*, vol. 30, no. 1, pp. 32–46, 1985.

[19] W. L. Thomas, "Television audience research technology, today's systems and tomorrow's challenges," *Consumer Electronics, IEEE Transactions on*, vol. 38, no. 3, pp. XXXIX–XXLII, 1992.

[20] A. L. Wang, "An industrial-strength audio search algorithm," in *ISMIR 2003*, 2003.

[21] B. Wild and K. Ramchandran, "Detecting primary receivers for cognitive radio applications," in *DySPAN 2005*, 2005, pp. 124 –130.

[22] Z. Zivkovic, "Improved adaptive Gaussian mixture model for background subtraction," in *ICPR 2004*, vol. 2. IEEE, 2004, pp. 28–31.