Camera Model Identification Using Audio and Visual Content from Videos

Ioannis Tsingalis, Christos Korgialas, and Constantine Kotropoulos

Department of Informatics, Aristotle University of Thessaloniki, Greece
The 2024 IARIA Annual Congress on Frontiers in Science, Technology, Services, and Applications

June 30 - July 04, Porto, Portugal

ARISTOTLE
UNIVERSITY OF
THESSALONIKI

Presenter's bio

Constantine Kotropoulos received the Diploma degree with honors in Electrical Engineering in 1988 and the PhD degree in Electrical & Computer Engineering in 1993, both from the Aristotle University of Thessaloniki.
He is currently a Full Professor in the Department of Informatics at the Aristotle University of Thessaloniki. He was a visiting research scholar in the Department of Electrical and Computer Engineering at the University of Delaware, USA during the academic year 2008-2009 and he conducted research in the Signal Processing Laboratory at Tampere University of Technology, Finland during the summer of 1993. He has co-authored 69 journal papers, and 222 conference papers, and contributed 9 chapters to edited books in his areas of expertise.
He is co-editor of the book "Nonlinear Model-Based Image/Video Processing and Analysis" (J. Wiley and Sons, 2001). His current research interests include forensics, audio, speech, and language processing, signal processing, pattern recognition, multimedia information retrieval, biometrics, and forensics.
Prof. Kotropoulos was a scholar of the State Scholarship Foundation of Greece and the Bodossaki Foundation. He is a senior member of the IEEE and a member of EURASIP, IAPR, and the Technical Chamber of Greece. He was a Senior Area Editor of the IEEE Signal Processing Letters and he has been a member of the Editorial Board of the journals: Advances in Multimedia, International Scholar Research Notices, Computer Methods in Biomechanics & Biomedical Engineering: Imaging & Visualization, Artificial Intelligence Review, MDPI Imaging, MDPI Signals, and MDPI Methods and Protocols.
Prof. Kotropoulos served as Track Chair for Signal Processing in the 6th Int. Symposium on Communications, Control, and Signal Processing, Athens, 2014; Program Co-Chair of the 4th Int. Workshop on Biometrics and Forensics, Limassol, Cyprus, 2016; Technical Program Chair of the XXV European Signal Processing Conf., Kos, Greece, 2017; Technical Program Chair of the 5th IEEE Global Conf. Signal and Information Processing, Montreal, Canada, 2017; General Chair of the 2022 IEEE 14th Image, Video, and Multidimensional Signal Processing Workshop, Nafplio, Greece; Technical Program Chair of the 2023 IEEE International Conf. on Acoustics, Speech, and Signal Processing, Rhodes, Greece.

Outline

- Camera Model Identification (CMI) is pivotal in multimedia forensic applications.
- The forensic analysis delves into various multimedia types, including audio recordings, images, and videos, to unravel the distinct signatures of different mobile phone brands/models.
- By exploiting these signatures, forensic analysts can accurately determine the particular device that recorded the multimedia content, providing crucial insights into various investigations, such as identifying the perpetrators behind a felony scene.

- The paper introduces a framework for CMI, treating it as a classification problem.
- Convolutional Neural Networks (CNNs) trained on either audio or visual content are employed for this purpose. Experimental findings showcase promising performance when employing either audio or visual content individually.
- Additionally, based on the classification decisions derived from the audio and visual content, late fusion is applied to these decisions using fundamental fusion rules, specifically the product and sum rules proposed by Kittler et al.[1]

[1] J. Kittler et al. "On combining classifiers". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 20.3 (1998), pp. 226–239.

Methodology

Dataset

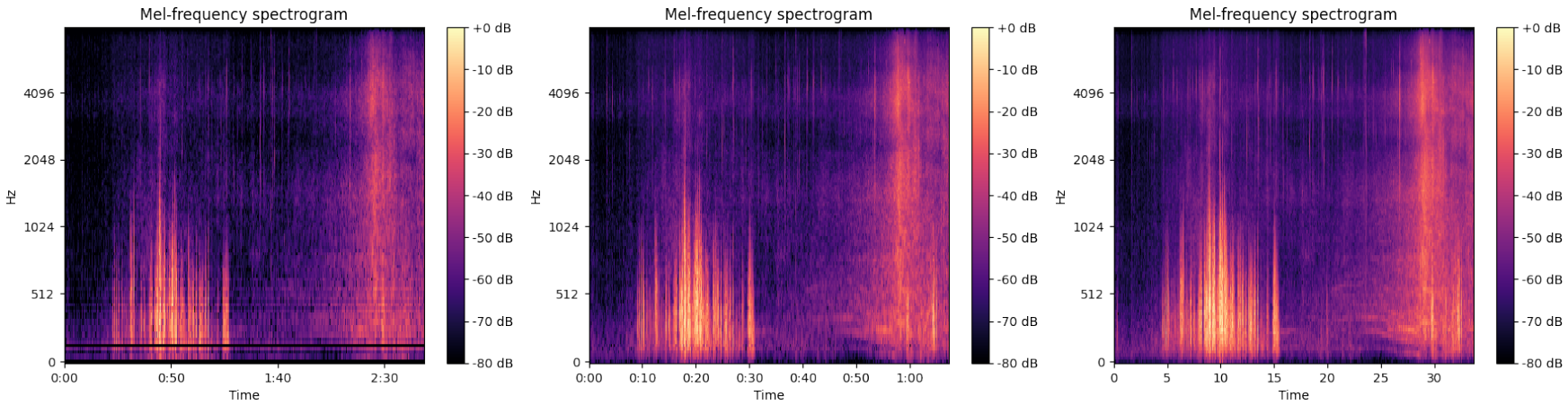
- The VISION dataset[2] is utilized, comprising images and videos captured across various scenes and imaging conditions.
  - 35 camera devices, representing 29 camera models and 11 camera brands, are encompassed within VISION.
- VISION includes 648 native videos, which remain unaltered post-capture by the camera.
  - These native videos were disseminated via social media platforms like YouTube and WhatsApp, with corresponding versions available in the dataset.
- The dataset is partitioned into training, testing, and validation sets to conduct a typical five-fold stratified cross-validation.

[2] D. Shullani et al. "Vision: A video and image dataset for source identification". In: *EURASIP Journal on Information Security* 2017 (2017), pp. 1–16.

Methodology

Content and Feature Extraction

- **Audio and Visual Content Extraction:**
  - *Audio content*: Extract audio content (.wav) from each video.
  - *Visual content*: Extract video frames from each video.
- **Audio and Visual Content Feature Extraction:**
  - *Audio content*: Three windows and hop parameters are used to construct a three-channel Log Mel-Spectogram for each audio recording.
  - *Visual content*: The raw video frames are used.



- We denote the instances of the mth modality as $\{\boldsymbol{\gamma}_m^{(n)}\}_{n=1}^{N}$.
- Let $\boldsymbol{W}_m^{[l]}$ and $f_{\boldsymbol{W}_m^{[l]}}^{[l]}$ be the parameters and the activation function of the lth layer on a Neural Network (NN) related to the mth modality, with $l = 1, \ldots, L$.
- Let $\{\boldsymbol{w}_m^{c',[L]}\}_{c'=1}^{C}$ be the collection of parameters belonging to the Lth layer where $\boldsymbol{w}_m^{c',[L]}$ is associated with the c'th classification node, with $c' = 1, \ldots, C$.

- Let

$$\Pr(\mathcal{C}_{c'} \mid \boldsymbol{\gamma}_m^{(n)} ; \boldsymbol{w}_m^{c',[L]}) = \frac{\exp(\boldsymbol{w}_m^{c',[L]^\top} \boldsymbol{a}_m^{(n)[L-1]})}{\sum_{c=1}^{C} \exp(\boldsymbol{w}_m^{c,[L]^\top} \boldsymbol{a}_m^{(n)[L-1]})} \tag{1}$$

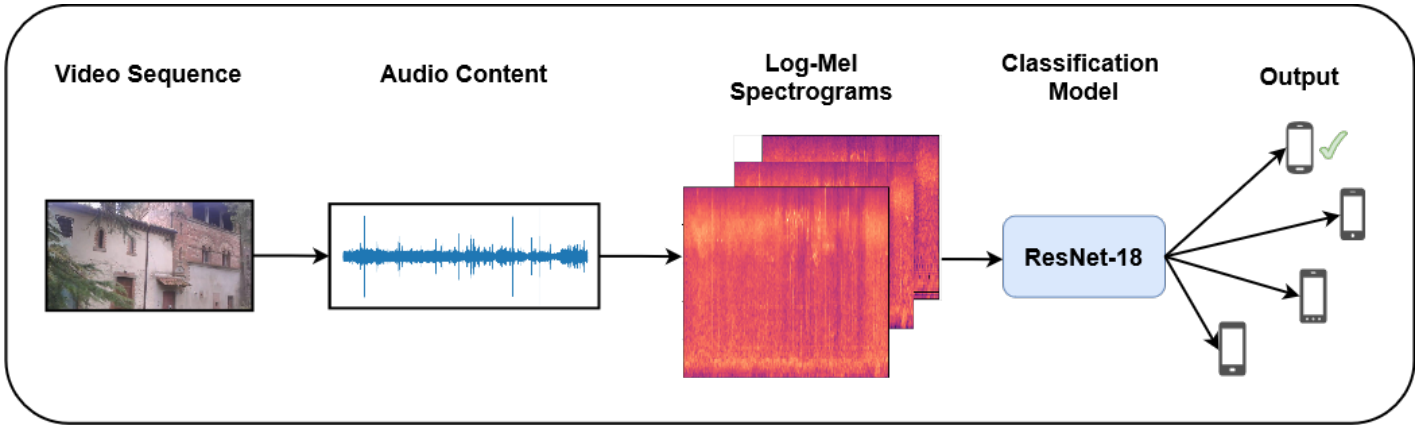
  be the classification probabilities of the c' classification node where

$$\boldsymbol{a}_m^{(n)[L-1]} = \left(f^{[L-1]}_{\boldsymbol{W}_m^{[L-1]}} \circ \cdots \circ f^{[1]}_{\boldsymbol{W}_m^{[1]}} \right) (\boldsymbol{\gamma}_m^{(n)}) \tag{2}$$

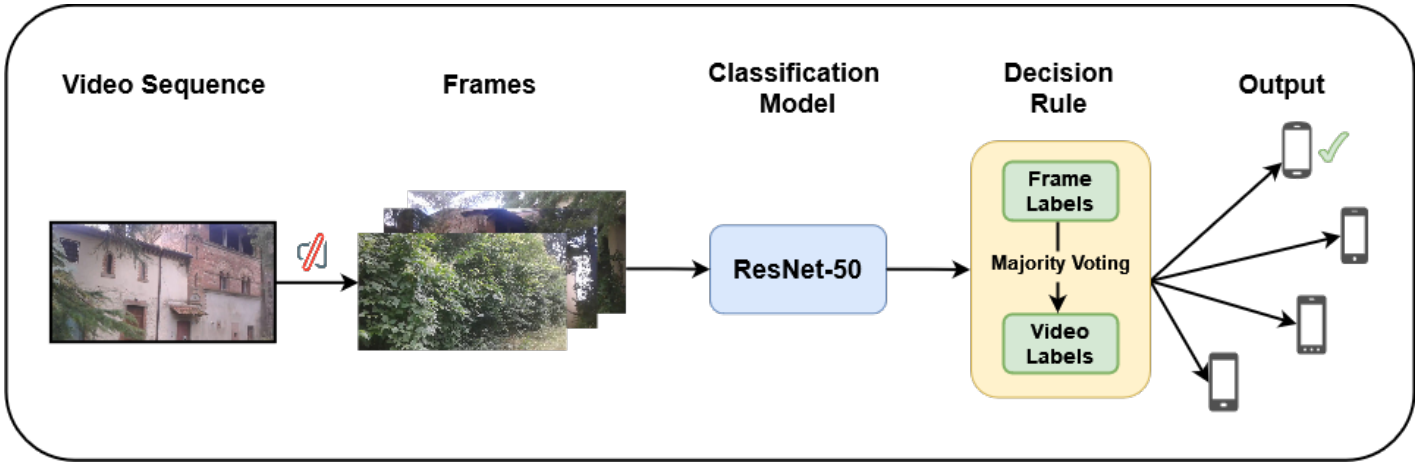
  is the activation output of the penultimate CNN layer, and $\circ$ denotes function composition.
- **Audio and Visual Content NN Classifier:**
  - Audio Content NN Classifier: `ResNet18`
  - Visual Content NN Classifier: `ResNet50`

Methodology

Classification Pipeline



(a) Audio classification pipeline using `ResNet18`.



(b) Video frame classification pipeline using `ResNet50`.

In addition, the classification probabilities of the nth sample $\boldsymbol{\gamma}_m^{(n)}$ related to the mth modality are given by

$$\boldsymbol{p}_m^{(n)[L]} = \begin{bmatrix} \Pr(\mathcal{C}_1 \mid \boldsymbol{\gamma}_m^{(n)} ; \boldsymbol{w}_m^{1,[L]}) \\ \Pr(\mathcal{C}_2 \mid \boldsymbol{\gamma}_m^{(n)} ; \boldsymbol{w}_m^{2,[L]}) \\ \vdots \\ \Pr(\mathcal{C}_C \mid \boldsymbol{\gamma}_m^{(n)} ; \boldsymbol{w}_m^{C,[L]}) \end{bmatrix} \in \mathbb{R}^C. \tag{3}$$

In the following, the superscript $[L]$ is omitted for simplicity. Given the samples $\{\boldsymbol{\gamma}_m^{(n)}\}_{n=1}^N$ of the mth modality, we obtain

$$\boldsymbol{P}_m = [\boldsymbol{p}_m^{(1)}, \boldsymbol{p}_m^{(2)}, \ldots, \boldsymbol{p}_m^{(N)}] \in \mathbb{R}^{C \times N}. \tag{4}$$

- **Unimodal Testing Procedure.** The predicted classes of each sample $\{\boldsymbol{\gamma}_m^{(n)}\}_{n=1}^{N}$ are given by

$$\boldsymbol{c}_m = [\mathcal{C}_m^1, \mathcal{C}_m^2, \ldots, \mathcal{C}_m^N]^\top \in \mathbb{R}^N, \tag{5}$$

where

$$\mathcal{C}_m^n = \underset{c=1,\ldots,C}{\arg\max} \left[\boldsymbol{p}_m^{(n)}\right]_c, \tag{6}$$

is the predicted class of the nth sample with $\mathcal{C}_m^n \in \{\mathcal{C}_{c=1}^C\}$.

- **Multimodal Testing Procedure.** The *product rule* is given by

$$\boldsymbol{P}_{\text{prod}} = \boldsymbol{P}_1 \odot \boldsymbol{P}_2 \odot \cdots \odot \boldsymbol{P}_M \in \mathbb{R}^{C \times N}, \tag{7}$$

where $\odot$ denotes the Hadamard, element-wise, product. The *sum rule* is given by

$$\boldsymbol{P}_{\text{sum}} = \boldsymbol{P}_1 + \boldsymbol{P}_2 + \cdots + \boldsymbol{P}_M \in \mathbb{R}^{C \times N}. \tag{8}$$

Experimental Evaluation

Unimodal Accuracy Results

Table 1: Accuracy (%) using visual and audio content

	Visual-ResNet-50			**Audio-ResNet-18**		
	Native	WhatsApp	YouTube	Native	WhatsApp	YouTube
Fold 0	88.31	67.53	77.02	96.10	93.50	91.9
Fold 1	85.70	83.11	72.97	94.80	90.90	93.24
Fold 2	89.60	63.63	77.02	90.90	88.31	95.94
Fold 3	89.47	68.42	63.51	93.42	94.73	82.43
Fold 4	88.15	64.47	78.37	94.73	88.15	95.94
Mean	88.24	69.43	71.77	93.99	91.11	91.89
$\pm$ StD	$\pm$ 1.4	$\pm$ 7.07	$\pm$ 5.44	$\pm$ 1.76	$\pm$ 2.66	$\pm$ 4.98

Experimental Evaluation

Late Fusion Accuracy Results

Table 2: Accuracy (%) using the product and sum rule

	Product Rule			**Sum Rule**		
	Native	WhatsApp	YouTube	Native	WhatsApp	YouTube
Fold 0	97.40	94.80	95.94	97.40	96.10	94.59
Fold 1	97.40	94.80	94.59	96.10	96.10	93.24
Fold 2	98.70	93.50	95.94	97.40	90.90	97.29
Fold 3	97.36	94.73	90.54	94.73	97.36	86.48
Fold 4	97.36	86.84	95.94	96.05	88.15	97.29
Mean	97.64	92.93	95.59	96.33	93.72	93.77
$\pm$ StD	$\pm$ 0.52	$\pm$ 3.08	$\pm$ 0.52	$\pm$ 0.99	$\pm$ 3.56	$\pm$ 3.97

Next, we study three null hypotheses:

- $H_{0,1}$: The classification performances achieved by the two fusion rules are equivalent.
- $H_{0,2}$: The classification performance achieved solely with visual content is equivalent to that achieved with the product rule.
- $H_{0,3}$: The classification performance achieved solely with audio content is equivalent to that achieved with the product rule.

Experimental Evaluation

Statistical Significance

- We have significant evidence against $H_{0,i}$, when the p-value falls within the range $[0.01, 0.05]$ for $i = 1, 2, 3$.
- We have highly significant evidence for $H_{0,i}$, when the p-value falls within the range $[0, 0.01]$, $i = 1, 2, 3$.
- When the p-value is greater that 0.05, we have not significant evidence against $H_{0,i}$ for $i = 1, 2, 3$.
- The p-values are computed by applying McNemar's significance test[3].

[3] Q. McNemar. "Note on the sampling error of the difference between correlated proportions or percentages". In: *Psychometrika* 12.2 (1947), pp. 153–157.

Experimental Evaluation

Statistical Significance

- Table 3 summarizes the p-values for $H_{0,1}$.
- Most of the p-values exceed the predetermined significance threshold, so we lack significant evidence against $H_{0,1}$.

Table 3: McNemar's p-values to evaluate the null hypothesis $H_{0,1}$

Folds	Native	WhatsApp	YouTube
Fold 0	0.0	1.0	1.0
Fold 1	1.0	1.0	1.0
Fold 2	1.0	0.5	1.0
Fold 3	0.5	0.5	0.3
Fold 4	1.0	1.0	1.0

Experimental Evaluation

Statistical Significance

- Table 4 summarizes the p-values for $H_{0,2}$ and $H_{0,3}$.
- It is evident that we have significant evidence against $H_{0,2}$.
- For $H_{0,3}$, most of the p-values exceed the predetermined significance threshold, so we lack significant evidence against $H_{0,3}$.

Table 4: McNemar's p-values to evaluate the null hypotheses $H_{0,2}$ and $H_{0,3}$

	Visual-ResNet-50			**Audio-ResNet-18**		
	Native	WhatsApp	YouTube	Native	WhatsApp	YouTube
Fold 0	0.023	10^{-5}	0.001	1.0	1.0	0.371
Fold 1	0.007	0.026	0.001	0.617	0.248	1.0
Fold 2	0.044	10^{-5}	0.001	0.041	0.133	0.479
Fold 3	0.041	10^{-5}	10^{-5}	0.248	0.617	0.007
Fold 4	0.045	10^{-4}	0.002	0.617	1.0	0.479

Conclusions

- A framework capable of device identification using audio, visual content, or a combination of both is introduced.
- CNNs are employed to address the device identification problem as a classification task.
- Experimental evaluation demonstrates a promising classification accuracy when independently using audio or visual content.
- Additionally, combining audio and visual content may lead to notable enhancements in classification performance, suggesting a potential area for further research.

Github Code

The code for the proposed framework can be found at:

`https://github.com/iTsingalis/IARIADevIDFusion`

Or by scanning the QR codes:

Acknowledgments

This research was supported by the Hellenic Foundation for Research and Innovation (HFRI) under the "2nd Call for HFRI Research Projects to support Faculty Members & Researchers" (Project Number: 3888).

Thank You!

Thank you very much for your attention.
Q & A?
Email costas@csd.auth.gr