Strategies for Countering Fake Information:

new trends in multimedia authenticity verification and source identification

Irene Amerini
amerini@diag.uniroma1.it

SAPIENZA
UNIVERSITÀ DI ROMA

About me

- Sep 2019: Dept. of Computer, Control, and Management Engineering "A. Ruberti"
  - Università degli studi di Roma "La Sapienza"

About me

- MFS-Lab, Media Integration and Communication Center (MICC),
  - Università degli Studi di Firenze, Italy

micc

- 2010 Scolarship, Digital Data Embedding Laboratory, Binghamton University, Binghamton (NY), US

BINGHAMTON
UNIVERSITY
STATE UNIVERSITY OF NEW YORK

- 2018 Visiting Fellow, Charles Sturt University, Wagga Wagga, Australia

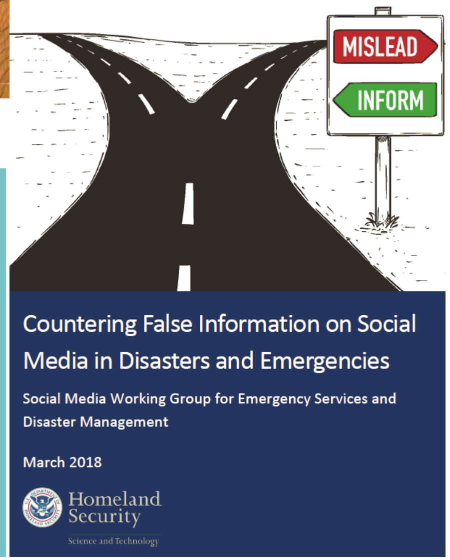
Charles Sturt University

ENDEAVOUR
Scholarships and Fellowships

Australian Government

The context: weaponized information

- The impact of Fake Information becomes fundamental, in the context of **weaponized information** and **information warfare**, where the organic propagation of virulent misinformation is under analysis.
    - Images and videos directly inoculate messages and have strong impact on personal opinions.

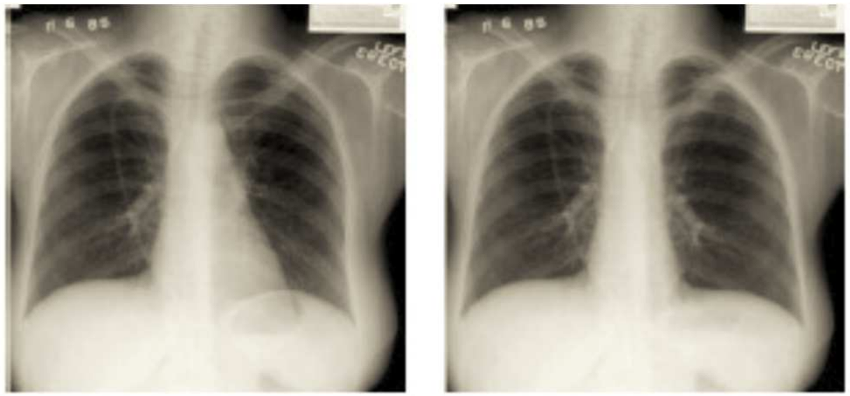


Propaganda/Military

In a court of law

Reputation attacks

Insurance frauds



Crime scene
alterations

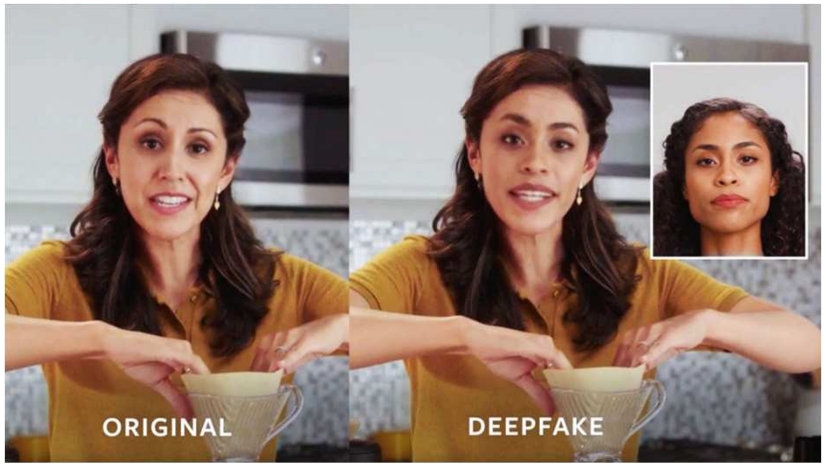
Not only images

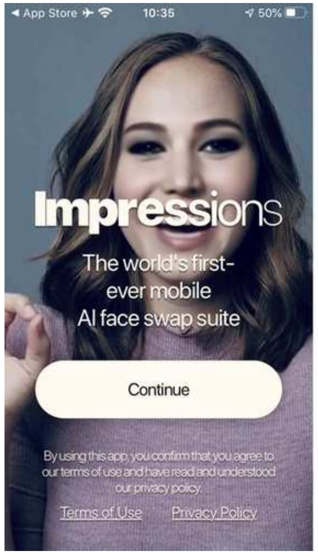
2016 2011

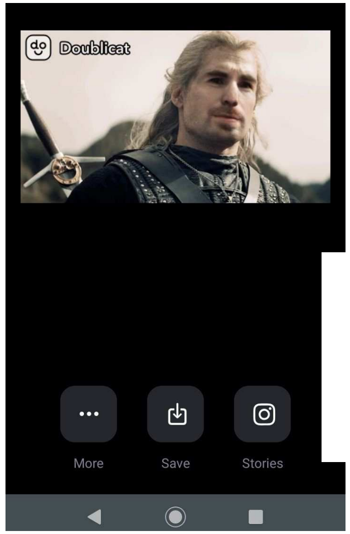

la Repubblica MILANO.it
Assalto alla stazione Centrale: treni pieni per le vacanze di Natale
Slideshow
1 di 20
Chiudi
Saggia Decisione
@saggiadecisione
Stazione di Milano Centrale ieri notte. Una vita in vacanza.
#coronavirus #Conte #Lombardia #COVID19italia #COVID—19 #StateACasa
11:10 AM · 8 mar 2020 · Twitter for iPhone
4 Retweet 4 Mi piace
(fotogramma)

Not only images

- Deep Fakes phenomena with AI
- Deepfake videos are AI-generated realistic sequences







https://beebom.com/best-deepfake-apps-websites/

On the web

- Tom Cruise (ago 2019)
  - https://www.youtube.com/watch?v=VWrhRBb-1Ig
- Matteo Renzi (sep 2019)
  - https://www.youtube.com/watch?v=E0CfdHG1sIs
- 20 celebrities (oct 2019)
  - https://www.youtube.com/watch?time_continue=37&v=5rPKeUXjEvE

How to «secure» an image or a video?

- Digital watermarking/Encryption
- Blockchain
- **Image and Video Forensics**

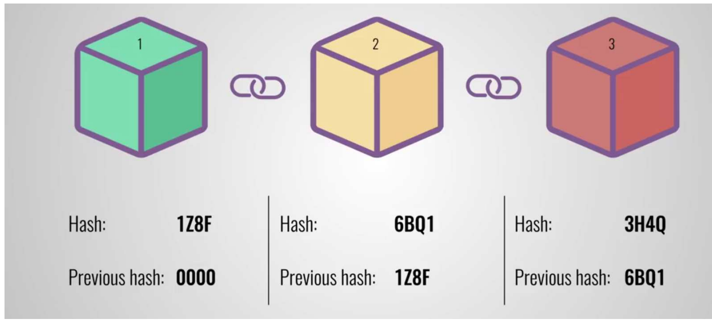
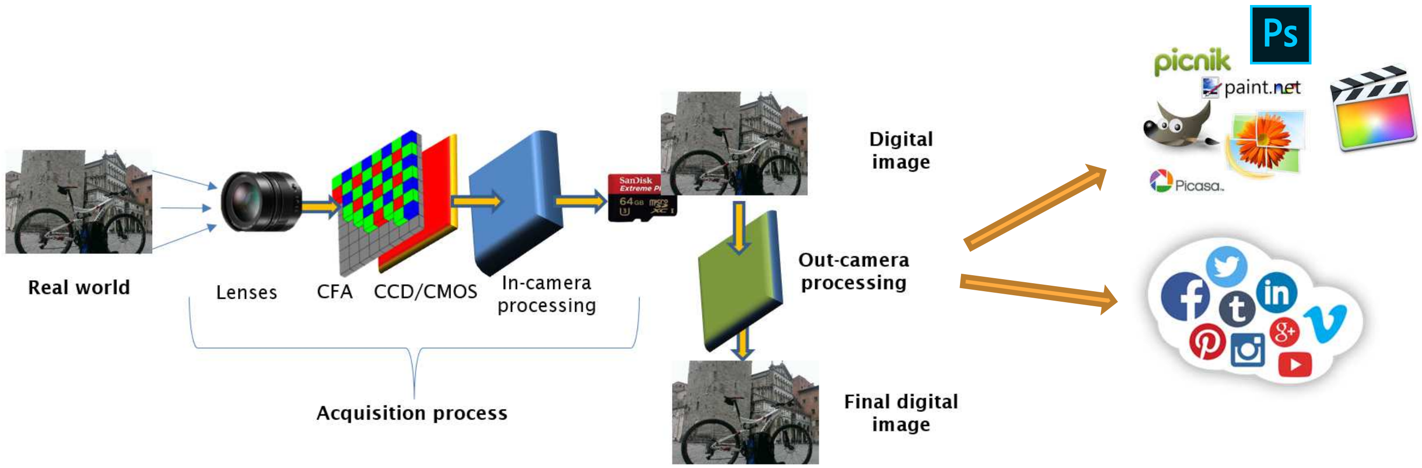


Image and Video Forensics

- To assess origin and originality of an image or video.
- Image and video forensic techniques gather information on the history of images and videos contents.
    - Each manipulation leaves on the media peculiar traces that can be exploited to make an assessment on the content itself.

Each phase leaves distinctive footprints!
- at the signal level
- at the metadata/file container level



Basic principles

- Only the image (video) and sometimes the device in our hands.
- No external information like metadata.

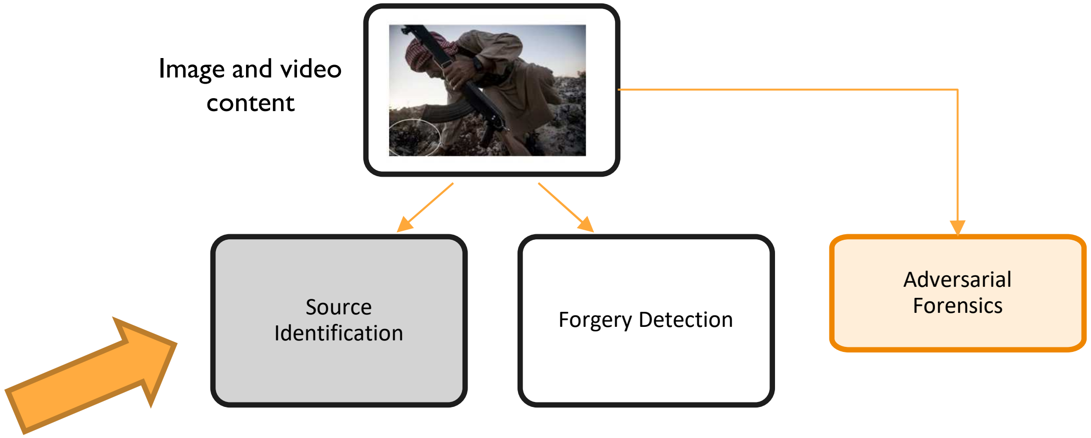
Blind: Original reference media is not required
- No side information like metadata

Passive:
Different from "active methods" which hide a mark in a picture when it is created like *digital watermarking*
- No specific on-device hardware required

- Acquisition process and post-processing operations leave a distinctive imprint on the data like a **digital fingerprint.**
  - *Fingerprint extraction*
  - *Fingerprint classification*

Image and Video Forensics

- **Source identification**: link a multimedia content to a particular (class of) acquisition device(s).
- **Forgery detection**: deciding on the integrity of the media
- **Adversarial forensics**/Counter forensics

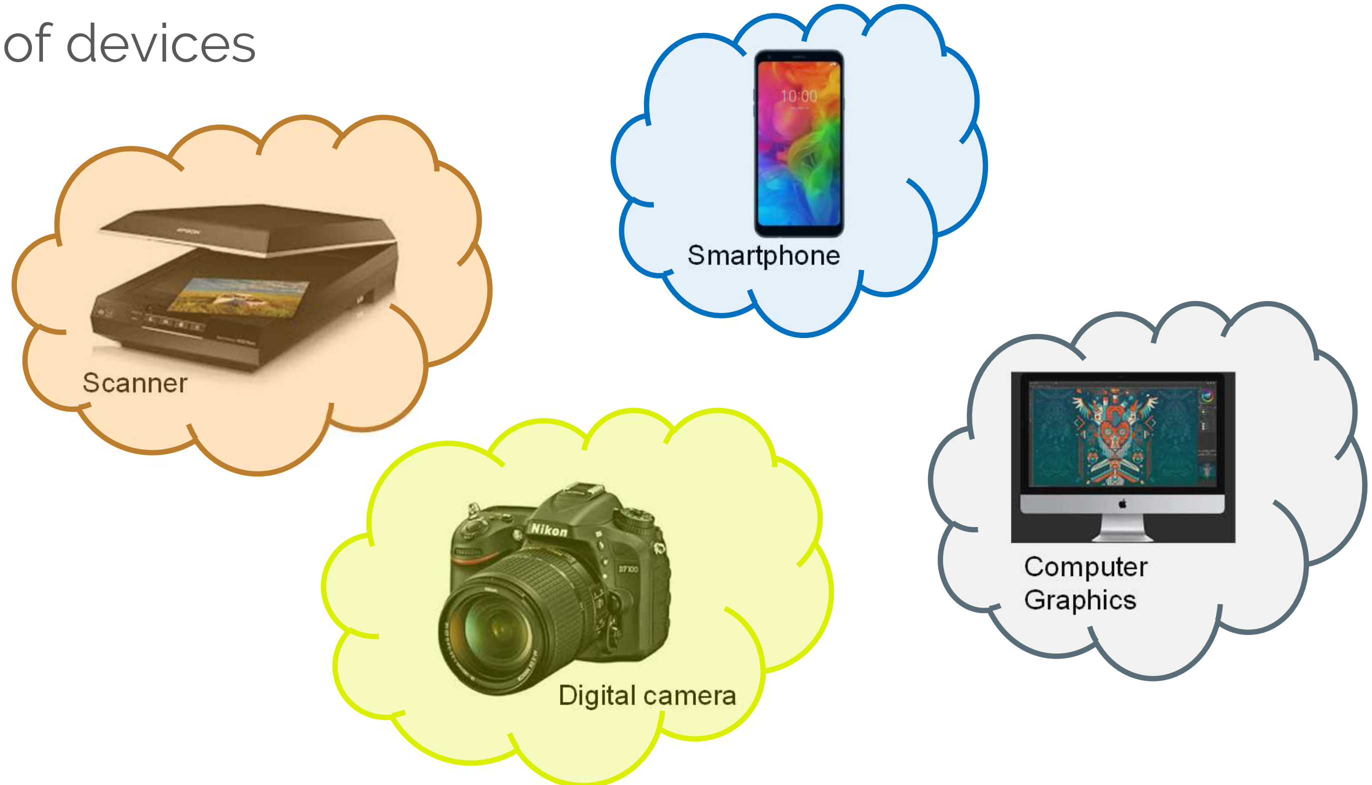


PART 1

Source Identification

Source identification

- Which **CLASS** of devices



Source identification

- Which **BRAND/MODEL**

Nikon?
Canon?
Sony?

Nikon D70
Nikon D3300

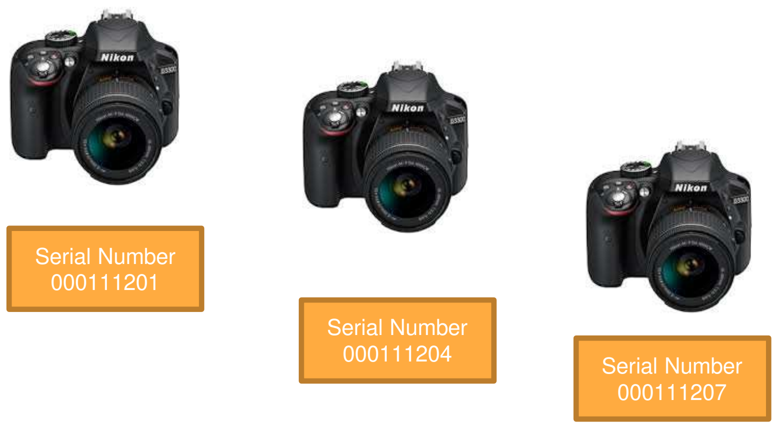
Sony cyber-shot dsc-h300
Sony a6000

Canon eos 1300d
Canon ixus 115 HS

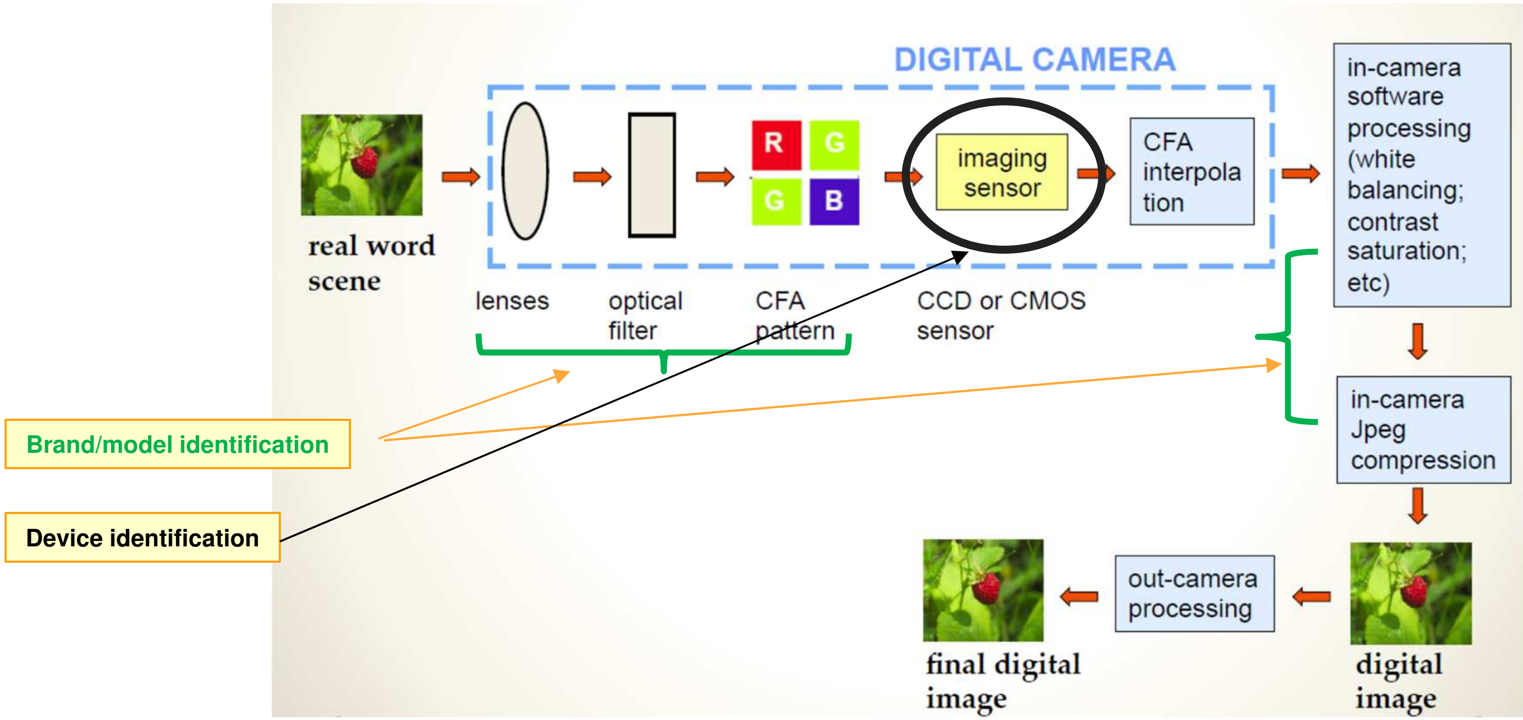
Source identification

- Which **DEVICE**

Which Nikon D3300?



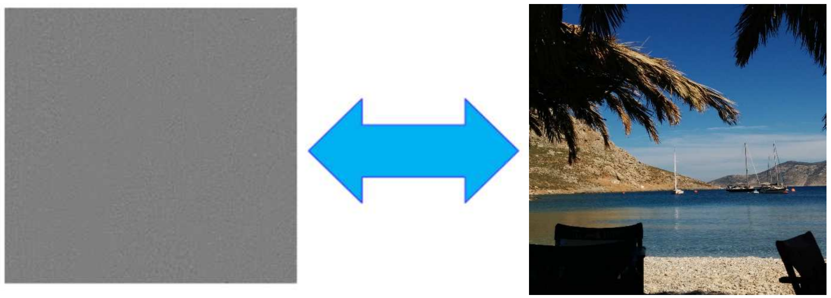
The acquisition process (in detail)



CCD sensor imperfections

- PRNU (Photo Response Non Uniformity Noise) is caused by **the different *sensitivity* of the sensors to light**
  - Due to the manufacturing process
  - Does not depend on temperature and time

- **If we capture this noise pattern, we can create a distinctive link between a camera and its photos**



[Fridrich at Al, TIFS 2006]

PRNU fingerprint model

A digital image **I** taken from camera C can be modeled as

$$\boldsymbol{I'} = \boldsymbol{I} + \boldsymbol{IK} + \boldsymbol{\theta}$$

Acquired image **Denoised image** **PRNU fingerprint** **Other noise terms (shot, readout etc..)**

$$\widehat{\boldsymbol{K}} = \frac{\sum_{i=1}^{N} \boldsymbol{W}_i \boldsymbol{I'}_i}{\sum_{i=1}^{N} (\boldsymbol{I'}_i)^2} \qquad \boldsymbol{W}_i = \boldsymbol{I'} - \boldsymbol{I}_F \qquad \boldsymbol{I}_F = \boldsymbol{F}(\boldsymbol{I'}) \approx \boldsymbol{I}$$

Observation: The PRNU pattern noise is a **multiplicative noise**

PRNU fingerprint detection

- Let **Y** be an input image (from the same camera C or another one)
- The presence of $\boldsymbol{K}$ in **Y** can be determined by means of the **correlation detector**

$$\boldsymbol{\rho = corr(W_Y, \hat{K}Y)}$$

Noise residual of image *Y*

Reference fingerprint

where:

$$corr(X,Y) = \frac{(X-\overline{X})\cdot(Y-\overline{Y})}{\|X-\overline{X}\|\ \|Y-\overline{Y}\|}$$

$$X \cdot Y = \sum_{i,j} X[i,j]Y[i,j]$$

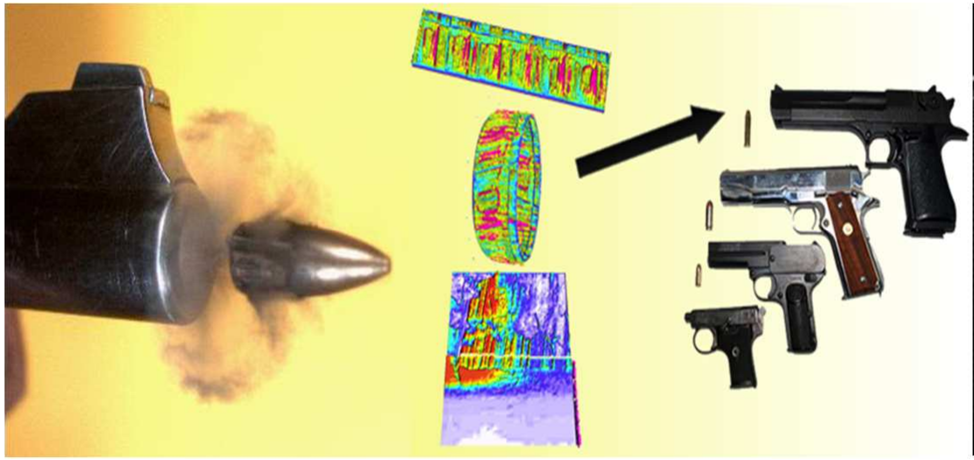
$$||X|| = \sqrt{X \cdot X}$$

High when **Y** was acquired by camera C with PRNU fingerprint **K**, low otherwise

A well known analogy

Firearms Identification

Digital Cameras Identification



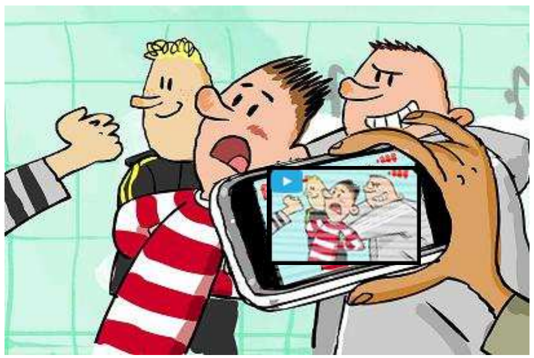
Social Network identification

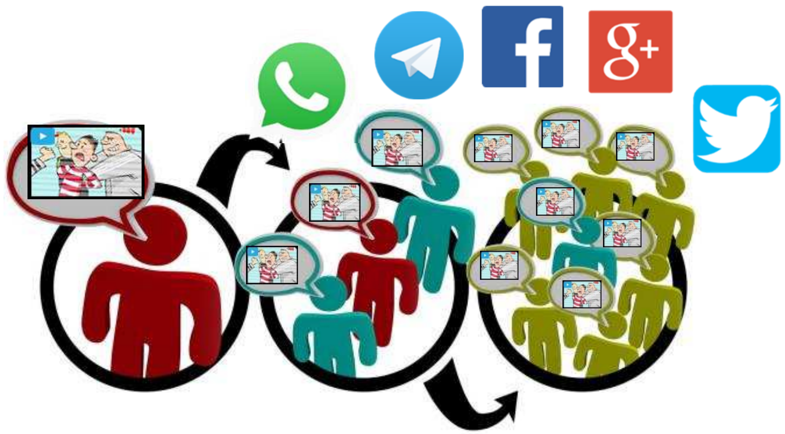
- In general, **source identification** is the process to link a multimedia content to a particular **acquisition device.**

- **Lately source identification also refers to establish the social network of origin.**

Social network identification

- Social Networks (SNs) are privileged channel for systematic and uncontrolled distribution of MM contents mainly images
  - Image shares are so quick that is not easy to follow their paths.
- In a forensic scenario (e.g. an investigation), it could be strategic understanding this flow so to reveal the intermediate steps followed by a certain content.
  - Resorting at the specific traces left by each SN on the image (**content** and **file**) due to the process each of them applies.





The rationale

- Uploading an image on a social network:
  - the process alters images
    - Resize, re-compression
    - New JPEG file structure
    - Rename
    - Meta-Data deletion/editing
  - each social network service (SNs) do different alterations with different rules

Some rules

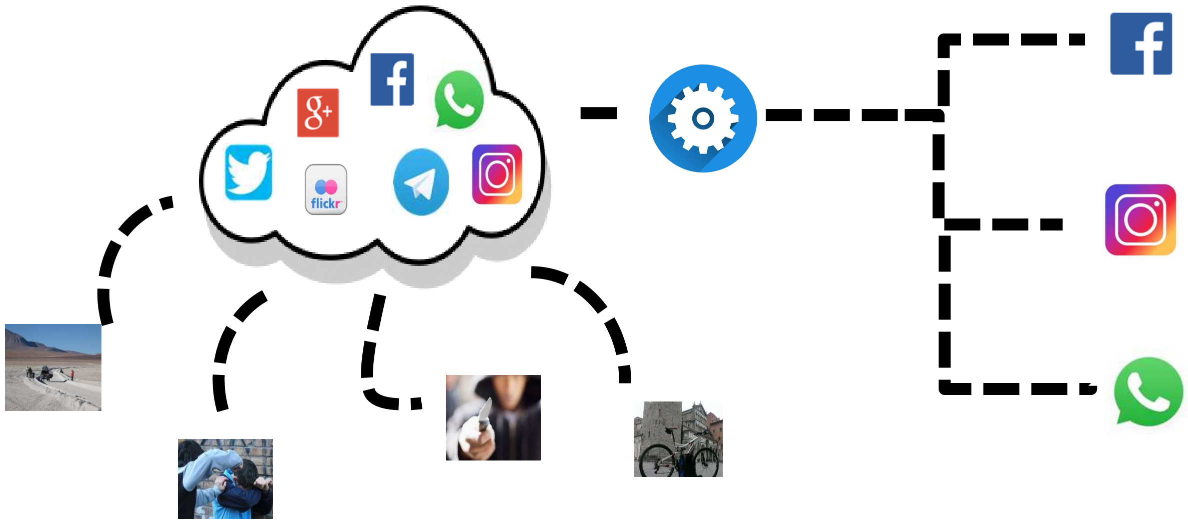
Social	EXIF		File Size		JPEG Compression	
	Camera Data	**Other Data**	**Resize**	**Resize Condition**	**Re-Compression**	**Re-Compression Condition**
Facebook	Delete	Delete	Yes	LQ: $M > 960$ HQ: $M > 2048$	Yes	Always
Google+	Maintain	Maintain/Edit	Yes	$M > 2048$	Yes	$M > 2048$
Flickr	Delete	Maintain/Edit	Yes	Depends on options	Yes	Depends on options
Tumblr	Maintain	Maintain/Edit	Yes	$M > 1280$	Yes	$M > 1280$
Imgur	Delete	Delete	No	Never	Yes	Image Size (MB) > 5.45 MB
Twitter	Delete	Delete	Yes	$M > 2048$	Yes	Always
whatsApp	Delete	Delete	Yes	$M > 1600$	Yes	Always
Tinypic	Maintain	Maintain/Edit	Yes	$M > 1600$	Yes	$M > 1600$
Instagram	Delete	Delete	Yes	$M > 1080$	Yes	Always
Telegram	Delete	Delete	Yes	$M > 2560$	Yes	Always

Social	Rename (image ID in bold)	Image Lookup	Other information
Facebook	11008414_**746657488782610**_8508378989307666639_n.jpg	YES	Upload resolution
Flickr	26742193671_**8a63f10c85**_h.jpg	YES	Download resolution (h=1600)
Tumblr	tumblr_**o3q9ghRCRh1vnf44lo9**_1280.jpg	YES	Download resolution (1280)
Imgur	04 - **Dw0KXG2**.jpg	YES	
Twitter	**CdqCPQ-WAAAzrHI.jpg**	YES	
WhatsApp	IMG-20160314-WA0038.jpg	NO	Receiving Date (2016-03-14)
Tinypic	1zqdirm.jpg	NO	
Instagram	1689555_**169215806798447**_744040439_n.jpg	YES	Upload Resolution
Telegram	422114602_5593965449613038107.jpg	NO	

The goal

Classify images according to the social network of provenance

- By identifying the distinctive and permanent trace "inevitably" imprinted in each digital content during the upload/download process by every specific social network.



The idea

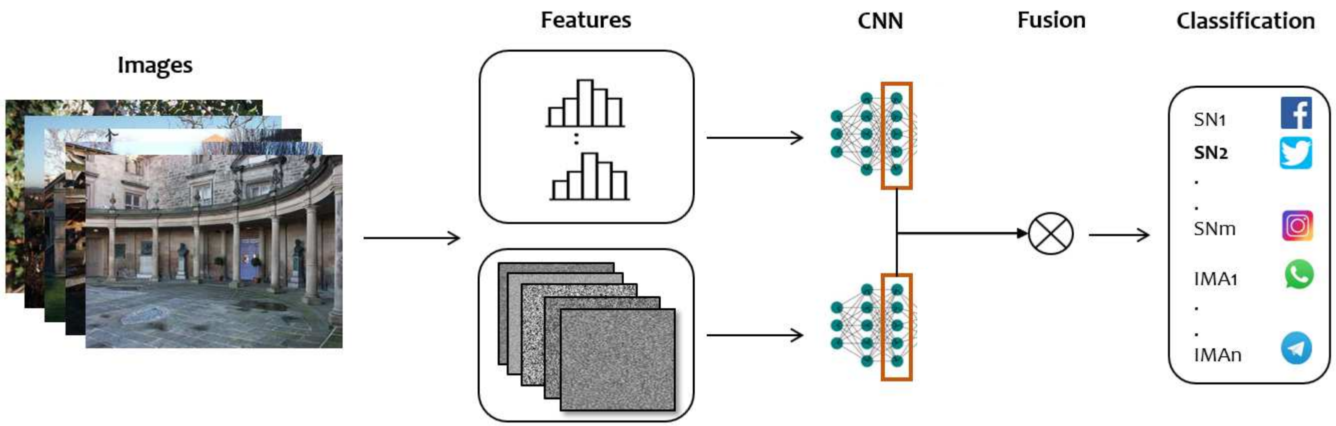
- Resorting at **image content-based features** to intercept processing affecting image itself such as JPEG multiple compressions, resizing, filtering and so on.

- Resorting at **metadata-based features** to take into account of changes to characteristics of the image file (e.g. quantization tables, image size).

Social Network Provenance: on image content

FusionNET: CNN-based framework for addressing the social network and instant messaging app identification

- Dual-modal features for image representation: the histogram of DCT and the sensor noise residuals
- Two CNN branches fed with the respective feature modalities to pull out activation vectors
- Fusion of activation vectors
- Classification of source SNs and IMAs of the images in question.



I. Amerini et Al, "Social Network Identification through Image Classification with CNN", IEEE Access 2019
I. Amerini et Al, "Image origin classification based on social network provenance", IEEE TIFS 2017

ENDEAVOUR Scholarships and Fellowships
Australian Government
Charles Sturt University

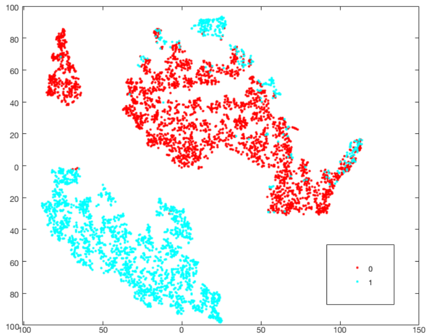
Some results

IPLAB dataset (4 devices, different resolutions, 7 SNs + original)

Classification (%) vs SNs	Facebook	Flickr	Google+	Instagram	Original	Telegram	Twitter	WhatsApp
Facebook	**91.31**	6.21	0.00	0.08	2.40	0.00	0.00	0.00
Flickr	0.90	**86.77**	0.03	0.18	3.26	0.70	8.14	0.02
Google+	0.01	0.03	**88.01**	0.48	11.44	0.02	0.00	0.02
Instagram	0.40	0.00	0.00	**98.80**	0.80	0.00	0.00	0.00
Original	0.00	0.00	0.00	0.00	**99.01**	0.99	0.00	0.00
Telegram	0.01	0.00	0.00	0.00	1.12	**98.87**	0.00	0.00
Twitter	0.11	2.00	0.00	0.00	1.51	0.11	**96.27**	0.00
WhatsApp	0.00	0.12	0.00	0.03	0.72	0.00	0.00	**99.13**

VISION dataset (10 smartphones, 3 SNs)

- t-SNE on VISION dataset: Facebook (class 0, red) and WhatsApp (class 1, cyan).



CNN	Fb	Wa	Orig
1D-CNN	97.76	98.61	99.99
2D-CNN	97.86	97.97	99.79
FusionNET	99.97	98.65	99.81

Single/Multiple shares

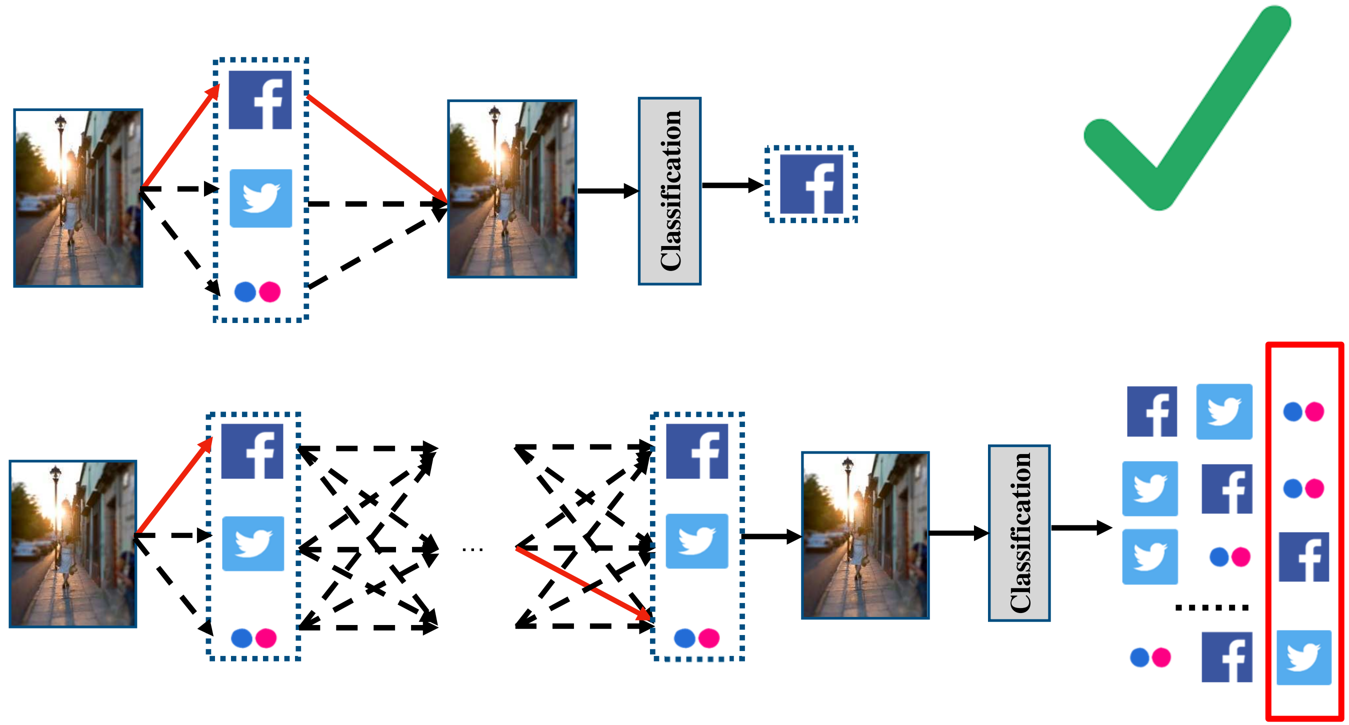
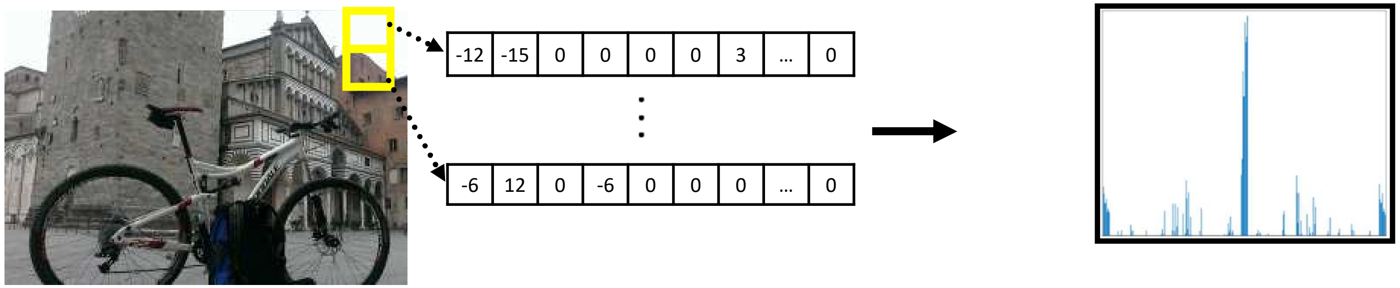


Image-based features



- BxB patches are considered (B=64)
- 8x8 block DCT coefficients are accumulated in histograms for **each of the 63 spatial frequencies** (DC is skipped!)
- Histograms are taken in a range of values between [-50, +50] [1] [2]
- A concatenated vector of 101 values is obtained for each DCT coefficient

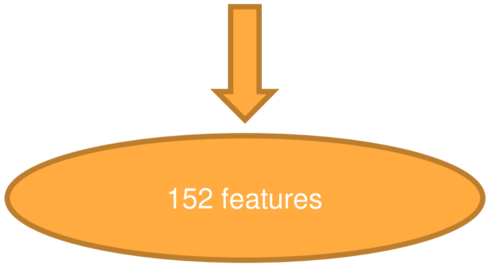
101x63 features

[1] Caldelli et al., *Image origin classification based on social network provenance*, TIFS 2017.
[2] Amerini et al., *Tracing images back to their social network of origin: A CNN-based approach*, WIFS 2017

Metadata-based features

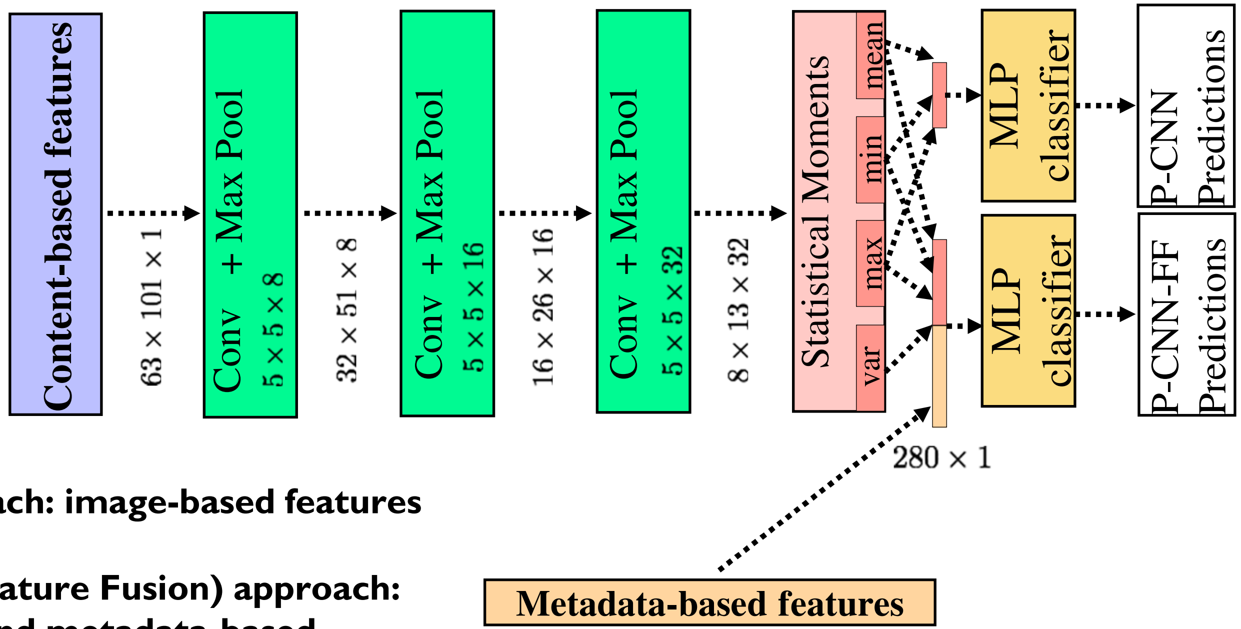
[3]

- Image dimensions (2 integers)
- Quantization tables (64x2=128 integers)
- Number of encoding tables used for AC & DC component (2 integers)
- Optimized coding and progressive mode (2 integers)
- Component information (18 integers)



[3] Q.-T. Phan et al. *Identifying Image Provenance: An Analysis of Mobile Instant Messaging Apps*. MMSP 2018.

Multiple up-down classification



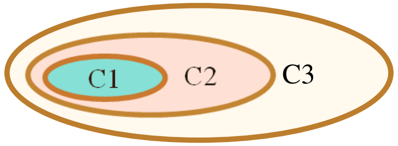
- **P-CNN approach: image-based features**

- **P-CNN-FF (Feature Fusion) approach: image-based and metadata-based features**

Q.-T. Phan, G. Boato, R. Caldelli, I. Amerini , "Tracking Multiple Image Sharing On Social Networks", IEEE International Conference on Acoustics, Speech, and Signal Processing, 2019.

Datasets (multiple)

- Three SNs have been considered: *Facebook*, *Twitter*, and *Flickr*.
- Up to **3 shares:**
  - **C1 (3 classes)**
  - **C2 (9 classes)**
  - **C3 (39 classes)**



$$\text{number of classes} = \sum_{k=1}^{K} (SN)^k$$

K= number of shares
SN= number of social networks

- **R-SMUD** (36000 images)
  - 50 raw images from *RAISE* [4] dataset
  - cropped top-left with 9:16 aspect ratio [377x600, 1012x1800, 1687x3000]
  - JPEG compressed using QF=50,60,70,80,90,100
- **V-SMUD** (20400 images)
  - 510 JPEG images selected from *VISION* [5] dataset (15 images x 34 cameras)

[4] D.-T. Dang-Nguyen, et al. *RAISE - A Raw Images Dataset for Digital Image Forensics*, ACM MM Systems, 2015.

[5] D. Shullani, et al. *VISION: a video and image dataset for source identification*, EURASIP JIS, 2017.

Exp. results: accuracy on single (C1) and double (C2) shares

Method	R-SMUD				V-SMUD			
	Patch level		*Image level*		*Patch level*		*Image level*	
	C1	C2	C1	C2	C1	C2	C1	C2
[11]	-	-	93.70	39.91	-	-	90.20	46.73
[12]	93.25	51.38	94.81	45.18	92.56	60.22	98.69	54.90
P-CNN	85.63	45.35	89.63	43.24	85.84	53.79	100.00	58.82
P-CNN-FF	**99.87**	**73.19**	**99.87**	**65.91**	**100.00**	**81.97**	**100.00**	**77.12**

- In the case of single share (3 classes), accuracy is satisfactory.
- In the case of double shares (9 classes), accuracy decreases but it is still good.

Exp. results (V-SMUD): double shares (C2)

True label \ Predicted label	FB	FL	TW	FB-FB	FB-FL	FB-TW	FL-FB	FL-FL	FL-TW	TW-FB	TW-FL	TW-TW
FB	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
FL	0.04	0.94	0.0	0.0	0.0	0.0	0.0	0.02	0.0	0.0	0.0	0.0
TW	0.12	0.0	0.0	0.0	0.0	0.63	0.0	0.0	0.1	0.0	0.0	0.16
FB-FB	0.75	0.0	0.0	0.2	0.0	0.0	0.02	0.0	0.0	0.04	0.0	0.0
FB-FL	0.27	0.0	0.0	0.0	0.73	0.0	0.0	0.0	0.0	0.0	0.0	0.0
FB-TW	0.12	0.0	0.0	0.0	0.0	0.59	0.0	0.0	0.18	0.0	0.0	0.12
FL-FB	0.06	0.0	0.0	0.0	0.0	0.0	0.88	0.06	0.0	0.0	0.0	0.0
FL-FL	0.0	0.06	0.0	0.0	0.04	0.0	0.08	0.82	0.0	0.0	0.0	0.0
FL-TW	0.04	0.0	0.0	0.0	0.0	0.29	0.0	0.0	0.53	0.0	0.0	0.14
TW-FB	0.0	0.0	0.0	0.0	0.0	0.0	0.02	0.0	0.0	0.98	0.0	0.0
TW-FL	0.06	0.61	0.0	0.0	0.0	0.0	0.0	0.08	0.02	0.0	0.24	0.0
TW-TW	0.12	0.0	0.0	0.0	0.0	0.63	0.0	0.0	0.1	0.0	0.0	0.16

True label \ Predicted label	FB	FL	TW	FB-FB	FB-FL	FB-TW	FL-FB	FL-FL	FL-TW	TW-FB	TW-FL	TW-TW
FB	0.98	0.0	0.0	0.0	0.0	0.0	0.02	0.0	0.0	0.0	0.0	0.0
FL	0.0	0.98	0.0	0.0	0.0	0.0	0.0	0.02	0.0	0.0	0.0	0.0
TW	0.0	0.0	0.22	0.0	0.0	0.35	0.0	0.0	0.35	0.0	0.0	0.08
FB-FB	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
FB-FL	0.0	0.0	0.0	0.0	0.96	0.0	0.0	0.04	0.0	0.0	0.0	0.0
FB-TW	0.0	0.0	0.18	0.0	0.0	0.39	0.0	0.0	0.39	0.0	0.0	0.04
FL-FB	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0
FL-FL	0.0	0.08	0.0	0.0	0.02	0.0	0.0	0.9	0.0	0.0	0.0	0.0
FL-TW	0.0	0.0	0.12	0.0	0.0	0.1	0.0	0.0	0.75	0.0	0.0	0.04
TW-FB	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0
TW-FL	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0
TW-TW	0.0	0.0	0.22	0.0	0.0	0.35	0.0	0.0	0.35	0.0	0.0	0.08

If we consider classification of «the last SN»: accuracy is **92% (P-CNN)** and **100% (P-CNN-FF)**.

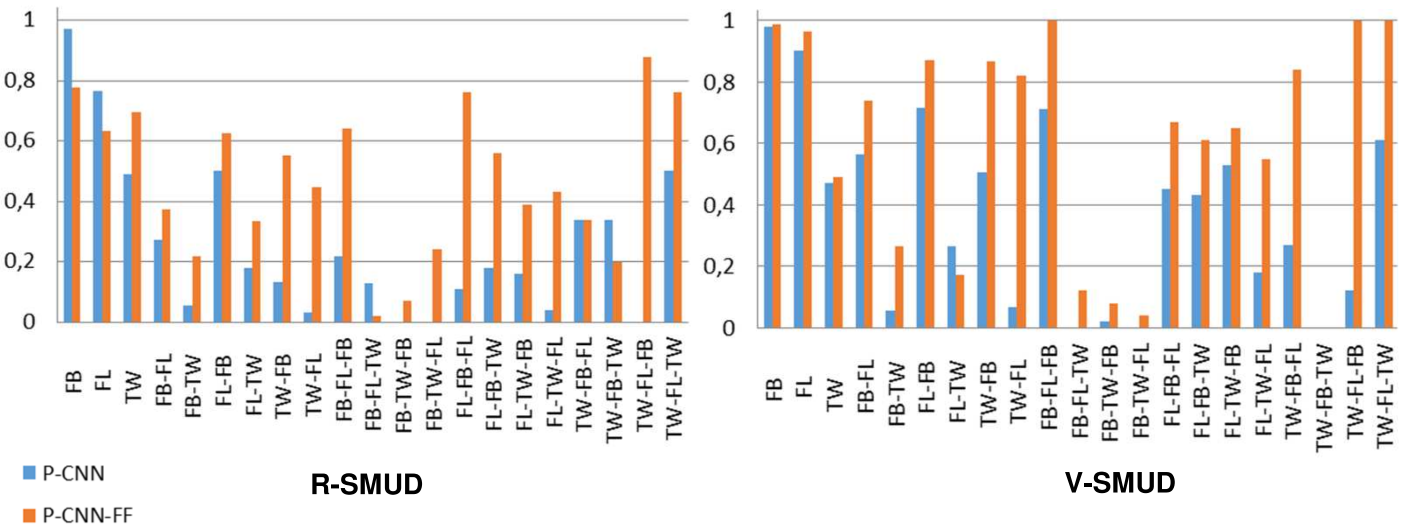
Exp. results: accuracy on triple shares (C3)

Consecutive up-downloads on the same SN do not affect the image, 39 classes are aggregated into 21 classes.

- FB-FB-FL → FB-FL
- FL-TW-TW → FL-TW
- ....

P-CNN-FF overall accuracy

- **aggregated case 60,6%**
- **only last SN 98,3%**



PART 2

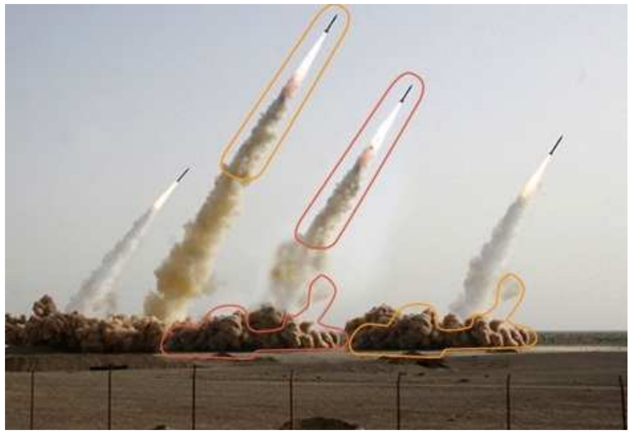
Authenticity verification

Kinds of manipulations

- Image manipulation categories:
  - Image Splicing
  - Copy-Move manipulation
  - Deepfakes

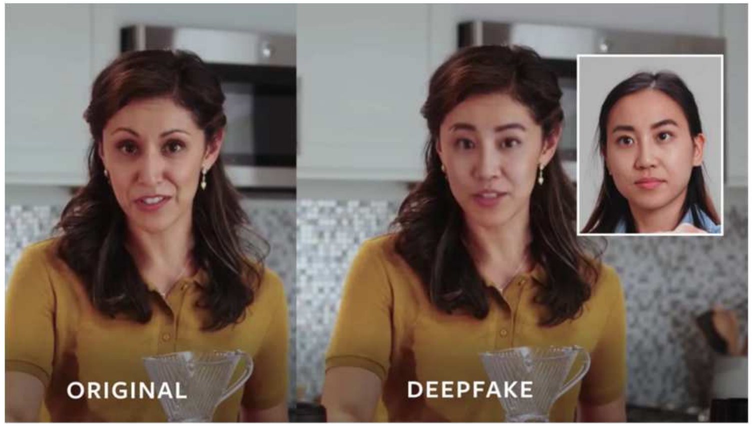
Kinds of manipulations

- Image manipulation categories:
  - Image splicing
  - Copy-Move manipulation
  - Deepfakes



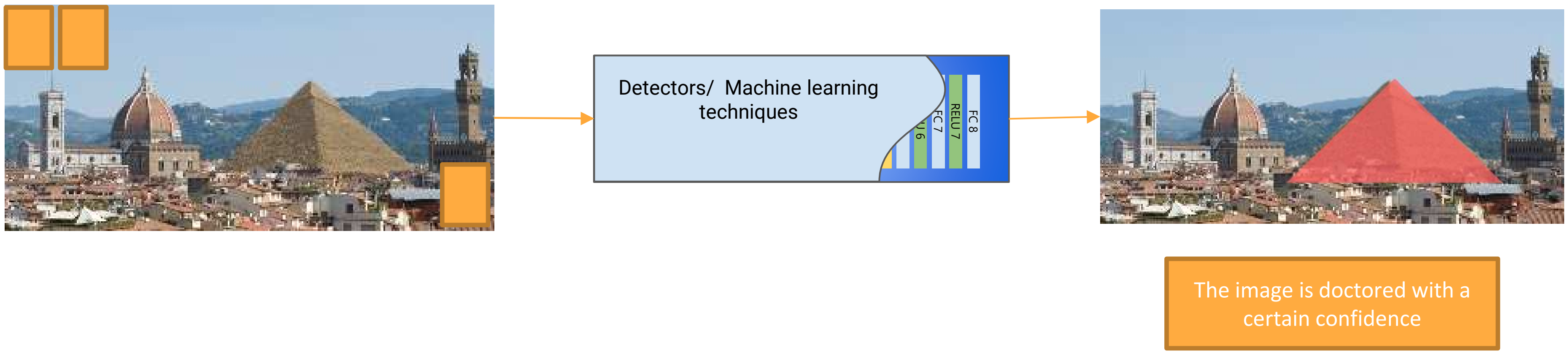
Kinds of manipulations

- Image manipulation categories:
  - Image splicing
  - Copy-Move manipulation
  - Deepfakes



Forgery detection

- **Research question:** how a doctored image/video be revealed and localized?
- Given a single probe image, detect if the probe was manipulated and provide mask(s)

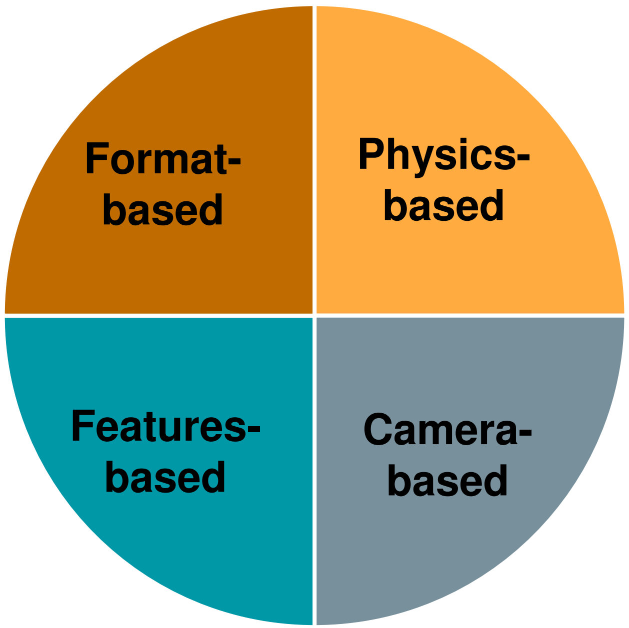


Techniques for tampering detection

- Techniques can roughly be separated in 4 categories

- Double JPEG compression
- First digits

- Illumination
- Trajectory

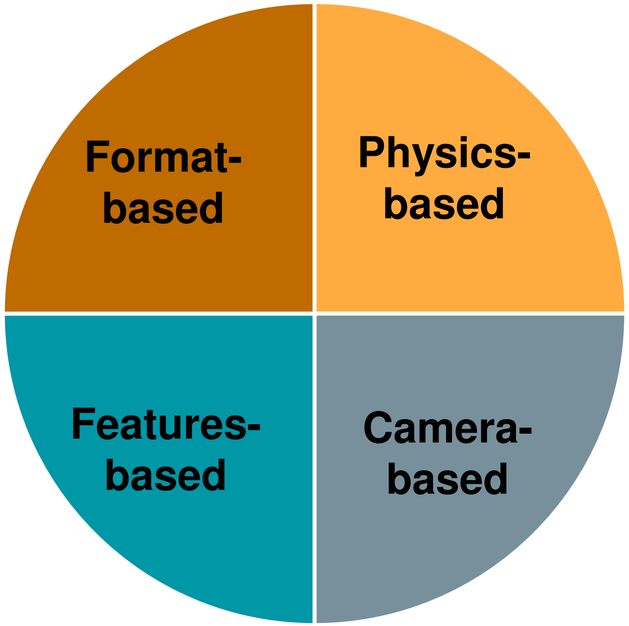


- SIFT

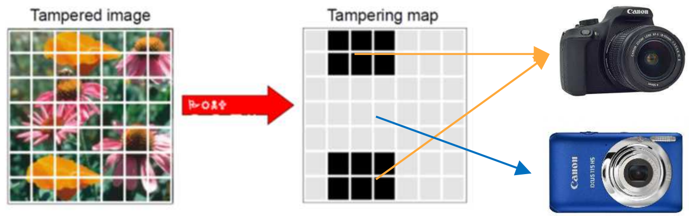
- PRNU

Techniques for tampering detection

- Techniques can roughly be separated in 4 categories

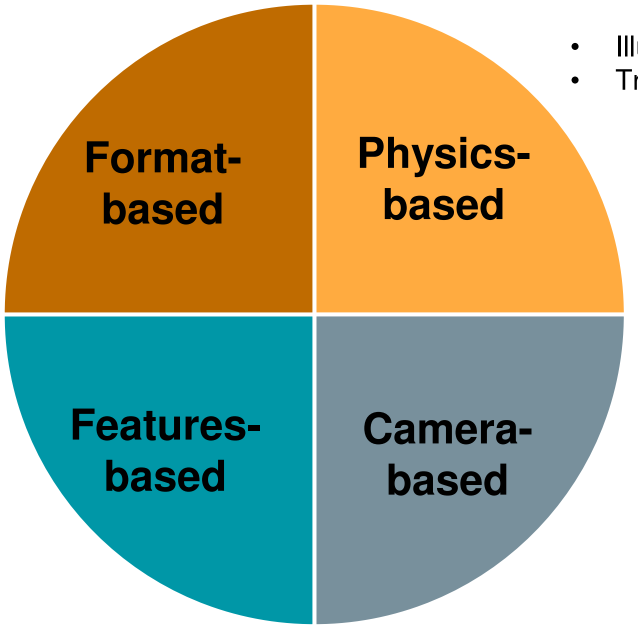


- PRNU

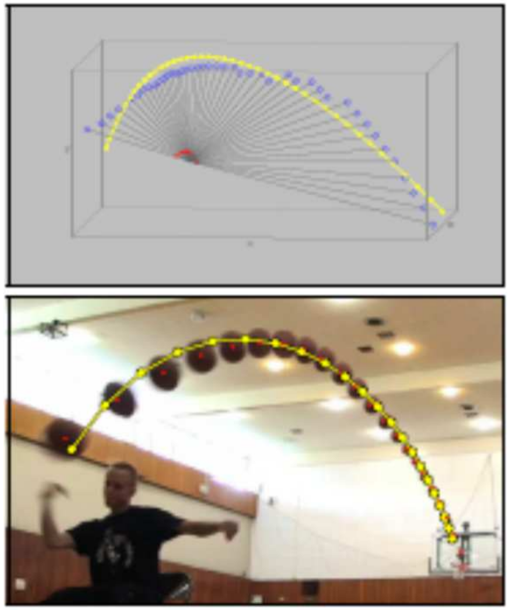


Techniques for tampering detection

- Techniques can roughly be separated in 4 categories



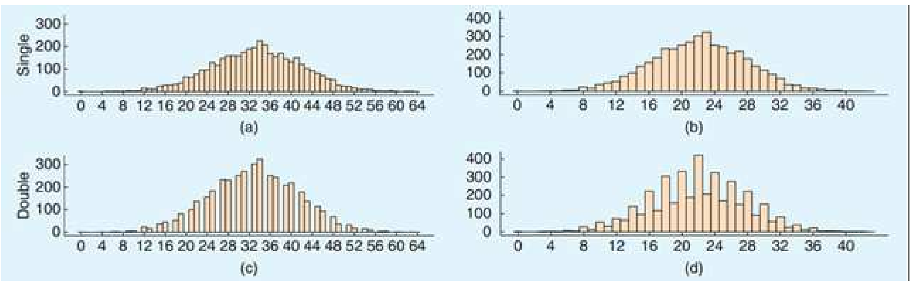
- Illumination
- Trajectory

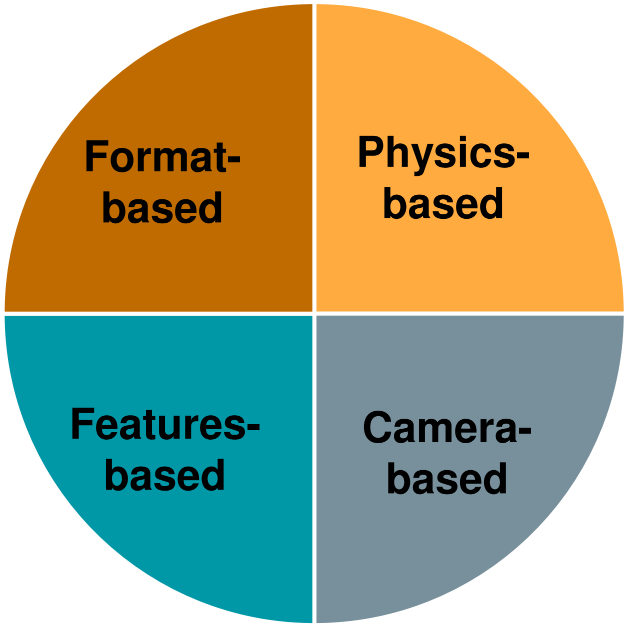


Techniques for tampering detection

- Techniques can roughly be separated in 4 categories

- Double JPEG compression
- First digits

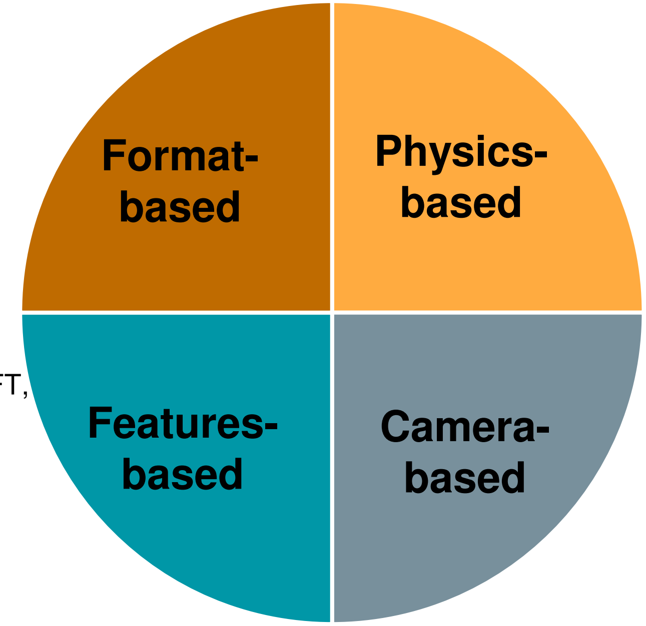




Techniques for tampering detection

- Techniques can roughly be separated in 4 categories

- Salient points (SIFT, SURF etc..)



Copy-move forgery detector (CMFD)

- A pioneer work to detect and localize «copy-move» image forgery
- It applies computer vision techniques to image forensics research problems
  - using local visual features and J-linkage clustering
- Definition of benchmarks datasets: MICC F220, MICC F2000, MICC-F600

I. Amerini, et Al, "A SIFT–based forensic method for copy-move attack detection and transformation recovery". IEEE Transactions on Information Forensics and Security, 2011

The copy-move manipulation

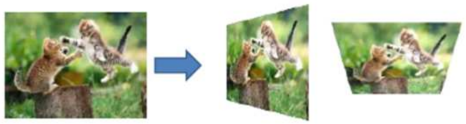
Hiding something

Duplicating something

Copy-Move Detection: salient point-based

When performing a cloning, usually a geometric transformation is applied to the cloned patch.

TARGET:
Forensic analysis should provide instruments to detect such a cloning and to estimate which transformation has been performed.



- In object detection and recognition, techniques based on scene modeling through a collection of salient points are well established.
- **SIFT** (*Scale Invariant Features Transform*) are usually adopted for their high performances and low complexity.

The proposed CMF detector

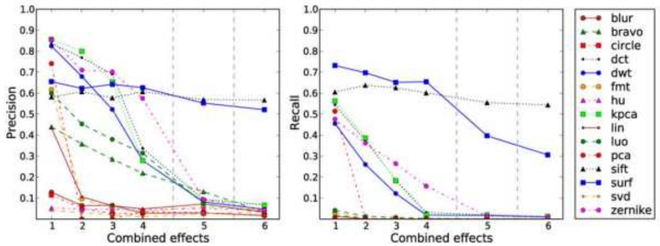
PHASE 1 PHASE 2 PHASE 3

Test image

SIFT feature extraction and matching

Clustering and forgery detection

Duplicated regions localization



Scaling, rotation, JPEG compression [Riess, TIFS'12]

Geometric transformation estimation

Correlation mask and segmentation

The syrian soldier case

CORRIERE DELLA SERA.it
Esteri

Home Opinioni Economia Cultura Spettacoli Cinema Sport Salute Tecnologia Scienze Motori Viaggi 27ora

Corriere della Sera › Esteri › *Il Pulitzer fa il tarocco, Ap lo licenzia in tronco* 23 gennaio 2014

GUAI PER IL FOTOREPORTER MESSICANO CONTRERAS

Il Pulitzer fa il tarocco, Ap lo licenzia in tronco

Manipolata un'immagine scattata in Siria

La foto «taroccata» (Ap)

Il peccato potrebbe apparire ai più veniale. Non all'Associated Press, una delle agenzie di stampa più prestigiose al mondo, che ha deciso di rompere i rapporti professionali con il fotoreporter messicano Narciso Contreras, già vincitore del Premio Pulitzer nel 2013, dopo aver scoperto che quest'ultimo ha manipolato un'immagine scattata durante la guerra civile in Siria. Il fotografo, utilizzando un moderno software, avrebbe fatto scomparire dallo scatto che immortalava un combattente siriano con il fucile in mano, la presenza, in basso a sinistra, di una telecamera di un collega.

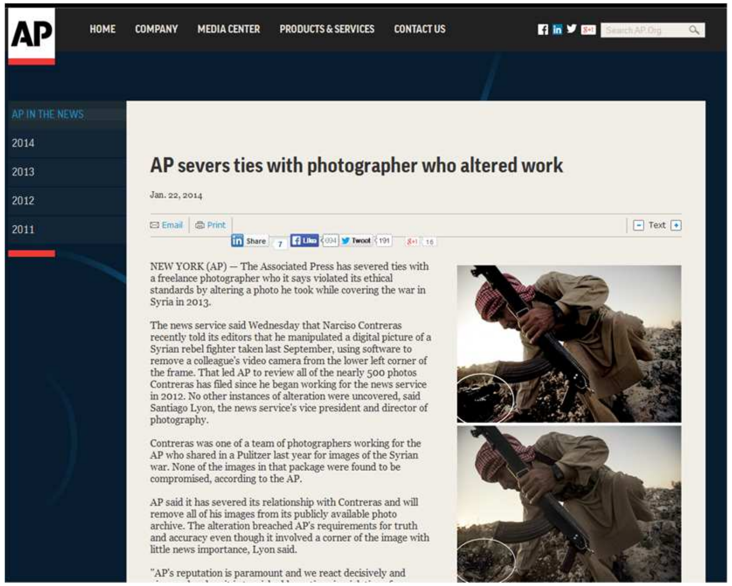
52% SODDISFATTO

OGGI IN esteri ›

La denuncia dell'Onu: «Le politiche del Vaticano hanno permesso abusi su bambini»

Giappone, le ultime lettere dei kamikaze «Mamma, ricorda che non ho pianto»

PIÙletti



AP

HOME COMPANY MEDIA CENTER PRODUCTS & SERVICES CONTACT US

AP IN THE NEWS
2014
2013
2012
2011

AP severs ties with photographer who altered work

Jan. 22, 2014

NEW YORK (AP) — The Associated Press has severed ties with a freelance photographer who it says violated its ethical standards by altering a photo he took while covering the war in Syria in 2013.

The news service said Wednesday that Narciso Contreras recently told its editors that he manipulated a digital picture of a Syrian rebel fighter taken last September, using software to remove a colleague's video camera from the lower left corner of the frame. That led AP to review all of the nearly 500 photos Contreras has filed since he began working for the news service in 2012. No other instances of alteration were uncovered, said Santiago Lyon, the news service's vice president and director of photography.

Contreras was one of a team of photographers working for the AP who shared in a Pulitzer last year for images of the Syrian war. None of the images in that package were found to be compromised, according to the AP.

AP said it has severed its relationship with Contreras and will remove all of his images from its publicly available photo archive. The alteration breached AP's requirements for truth and accuracy even though it involved a corner of the image with little news importance, Lyon said.

"AP's reputation is paramount and we react decisively and

VENERDÌ 24 GENNAIO 2014, AGGIORNATO ALLE 17:53

CORRIERE FIORENTINO.it

«La foto? Si (s)trucca così»

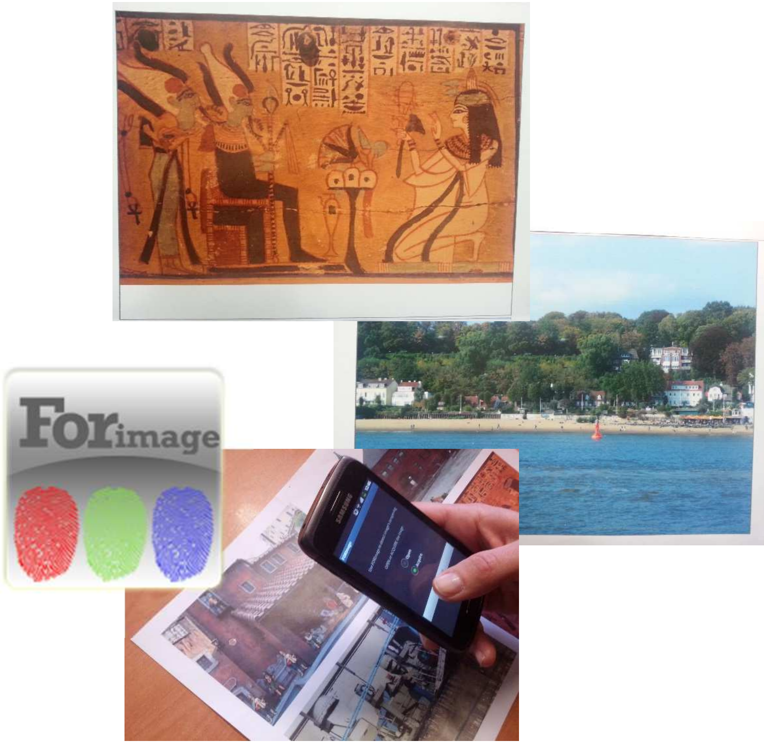
CRONACHE | Roberto Caldelli, dell'Università di Firenze, ha messo a punto il software che ha smascherato il fotoreporter dell'Ap *di L. Rebecchi*

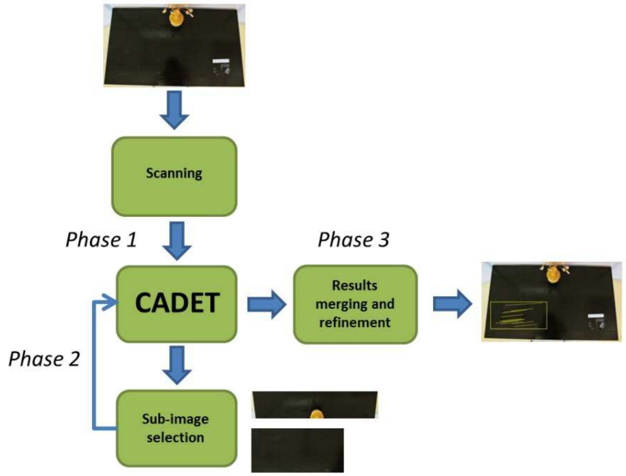
la Repubblica.it Foto del reporter licenziato da Ap: l'analisi del ritocco

R.it Tecnologia

"Ecco come si trucca un'immagine": l'esperto spiega il fotoritocco del Pulitzer

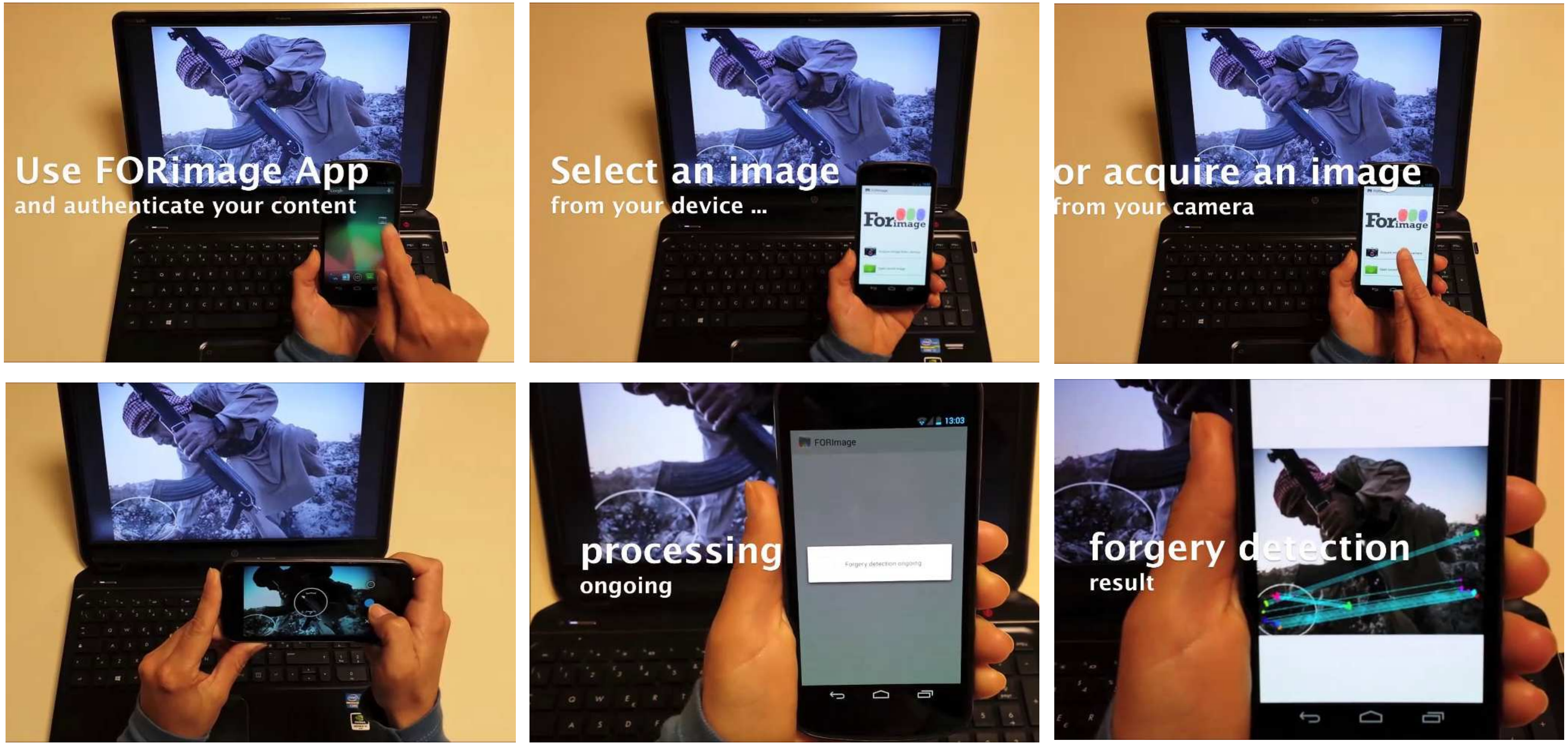
Printed images





I. Amerini, R. Caldelli, A. Del Bimbo, A. Di Fuccia, A. P. Rizzo, L. Saravo, "Detection of manipulations on printed images to address crime scene analysis: A case study", Forensic Science International, 2015.

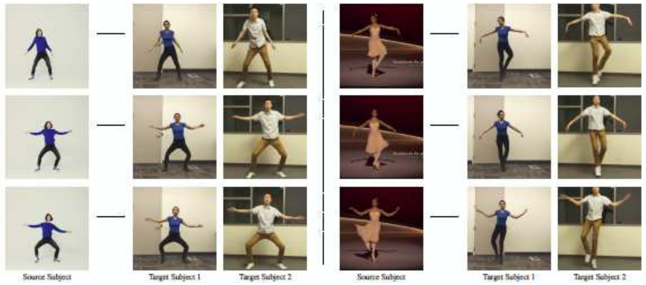
FORimage app



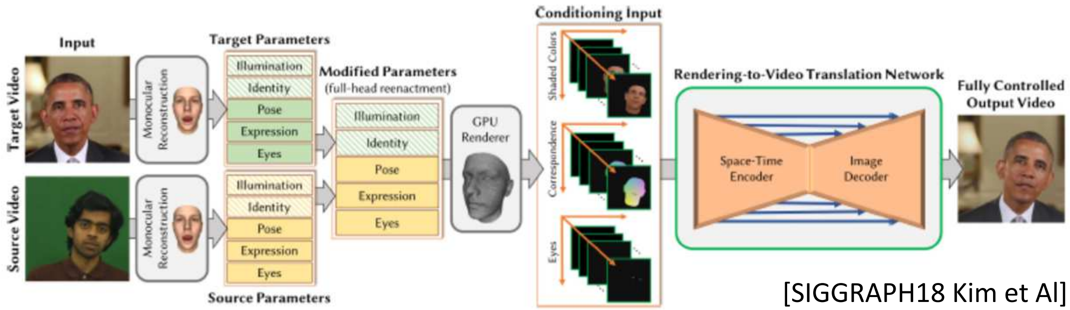
Deepfake phenomena with AI

- Many techniques: FaceTransfer, Face2Face, DeepFake, Deep Video Portaits, FaceSwap etc..

reddit



Everybody can dance now
[Chan, Efros 2018]



[SIGGRAPH18 Kim et Al]

Facial video editing

- Face Swap vs Reenactment/ Video graphics vs Deep Learning (GAN)

[Niesser, CVPR2016]

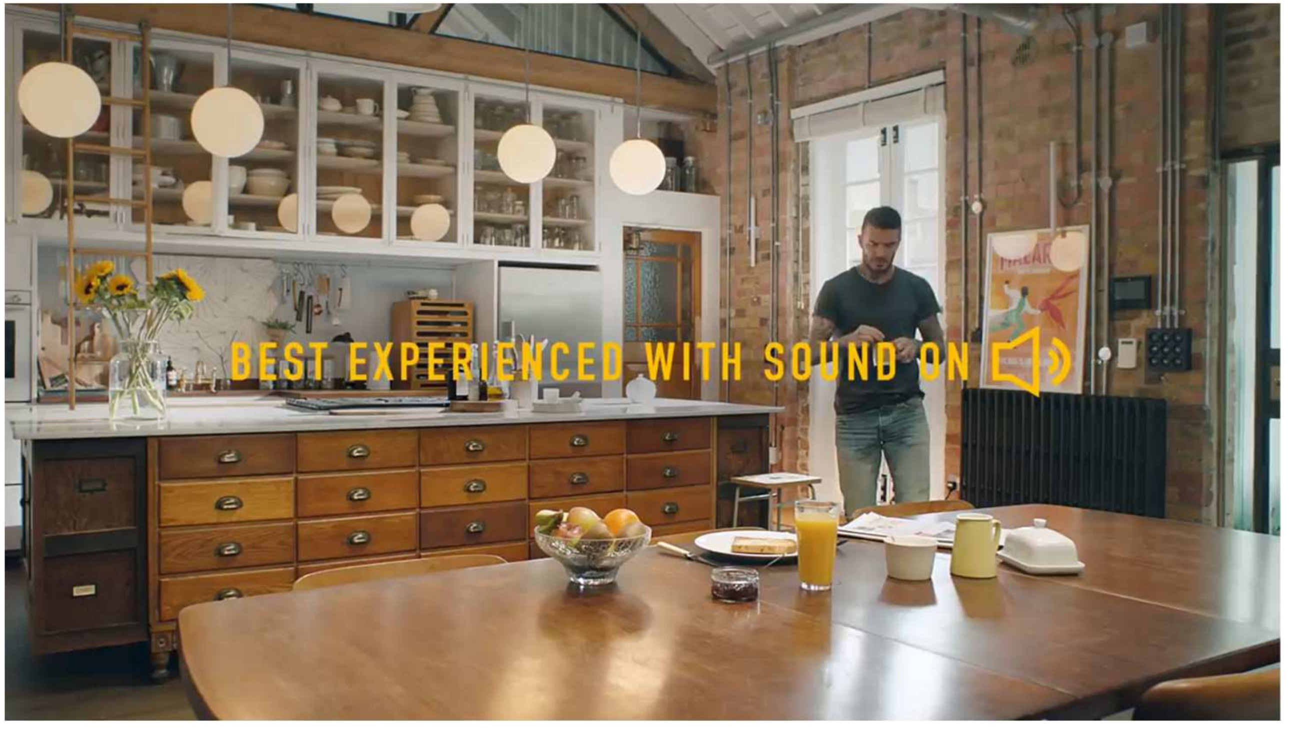
[FaceSwap]

[FakeApp, Reddit]

What Obama is saying?

https://www.youtube.com/watch?v=cQ54GDm1eL0

Synthesia dubbing and storytelling

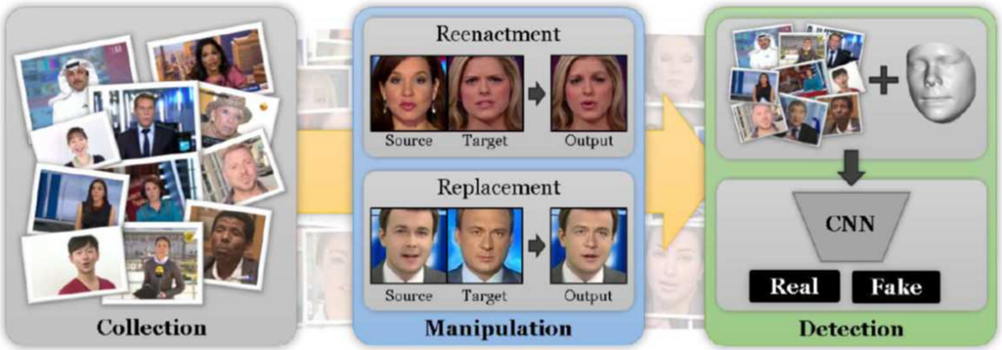


A proliferation of datasets

- FaceForensics dataset: Video Dataset for Forgery Detection in Human Faces generated with the F2F facial reenactment algorithm altering facial expressions with the help of a reference actor
- FaceForensics++ (F2F, FaceSwap, DeepFake, Neural Textures) 1000 images for each manipulation methods
- Google
- Facebook
- ….

Learning to Detect Manipulated Facial Images

- Face tracking method: extract the region of the image covered by the face; this region is fed into a learned classification network that outputs the prediction (RGB patch).
- Classification based on XceptionNet [13] outperforms all other variants in detecting fakes.
- Evaluation of different state-of-the-art classification methods.

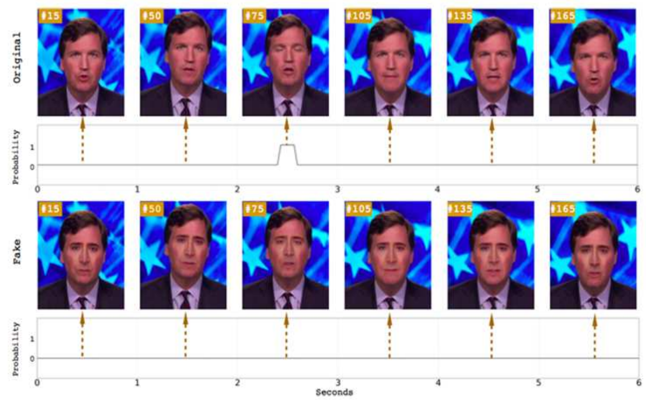


[Rossler et Al, ICCV 2019]

Deepfake videos detection in literature

Deepfake videos are usually detected by resorting at **frame-based** approaches which look for:

- *spatial inconsistencies in frames*
- *semantic anomalies (e.g. different colour of the eyes)*
- *eye blinking absence*
- *biological signal*
- *symmetry inconsistencies*



Our approach

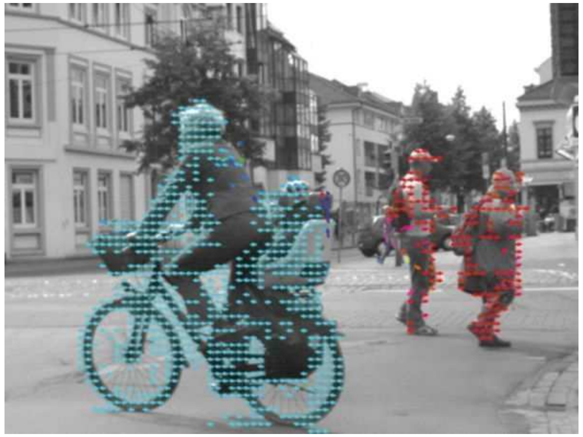
A **sequence-based** approach is introduced by looking at possible dissimilarities in the video temporal structure

- *Optical flow fields* have been extracted from the video sequence
- Motion vectors should exploit different inter-frame correlations between fake and original videos
- Such an information is used as input of CNN-based classifiers.

[Amerini et Al, “Deepfake Video Detection through Optical Flow based CNN”, Human Behaviour and Understanding Workshop, ICCV 2019]

The optical flow field

- **Optical Flow fields** describe the apparent motion of objects in a scene due to the relative motion between the observer (the camera) and the scene itself.



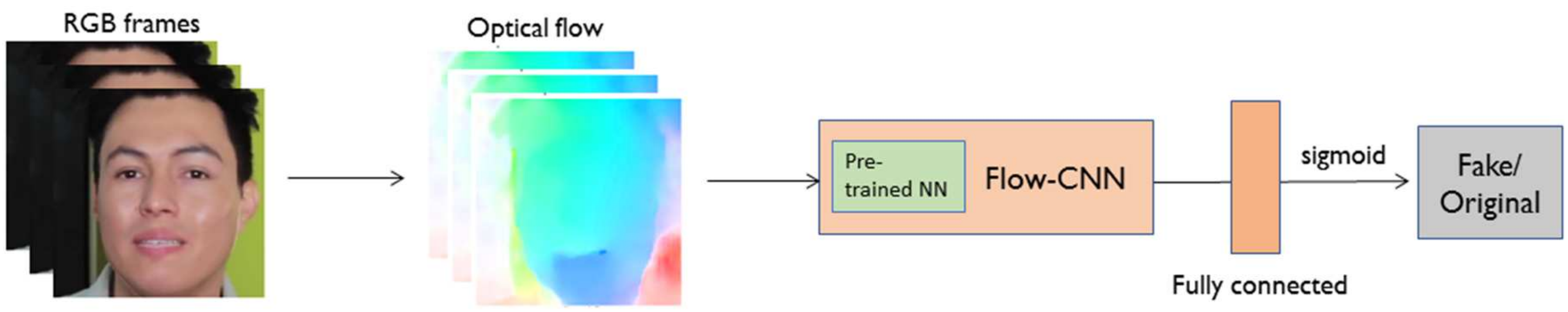
- Given two consecutive frames *f(t)* and *f(t+1)*:

$$f(x, y, t) = f(x+\boldsymbol{\Delta x}, y+\boldsymbol{\Delta y}, t+1)$$

  - OF fields, in our experiments, have been computed by resorting at PWC-Net.

The proposed pipeline

- OF fields are used as input of a semi-trainable neural network
- Neural networks such as *VGG-16* or *ResNet50*, pre-trained on Optical Flow, have been tested
- The last convolutional layers and the dense ones are trained on deepfake dataset



Test set-up

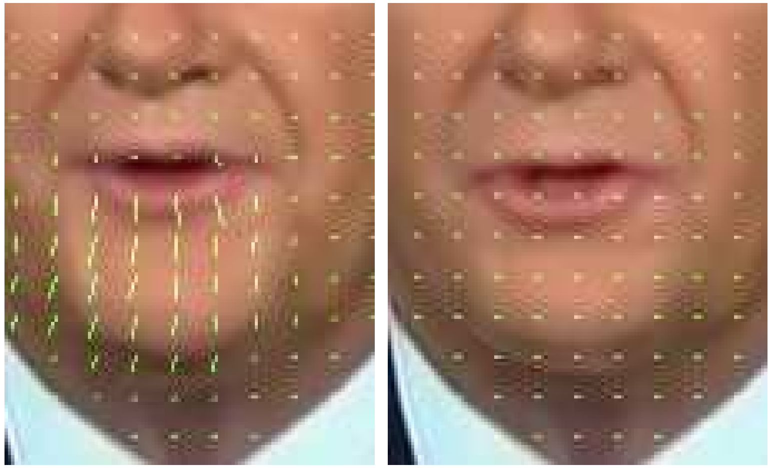
Dataset

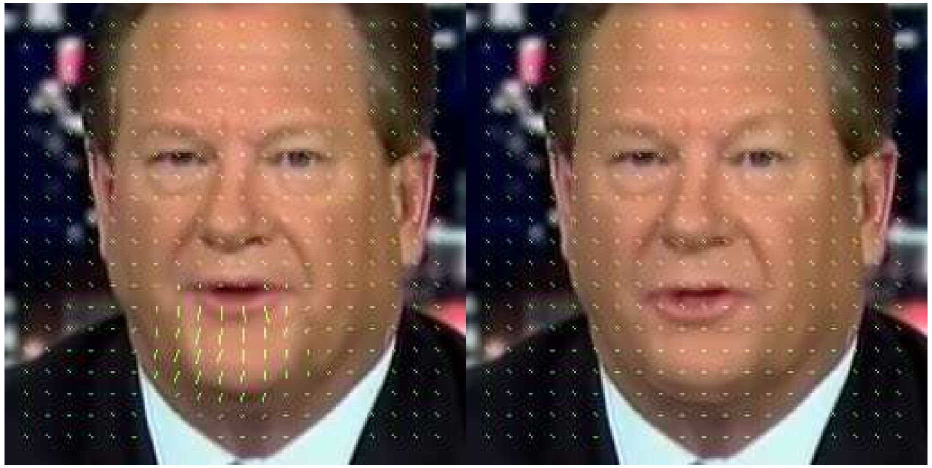
- *FaceForensics++*
- 1000 videos (original and fake for each kind of manipulation)
- 720 for training set, 140 for validation and 140 for test set

- A patch of 300x300 pixels, around the face, is cropped from each frame
- A squared patch of 224x224 pixels is randomly chosen and flipped left-right for data augmentation
- Adam optimizer with learning rate 10^{-4}, default momentum values and batch size of 256 is used.

Experimental results

- Looking at MVs, particularly around the mouth, a different distribution of the OF field is appreciable:
    - Deepfake case is smoother

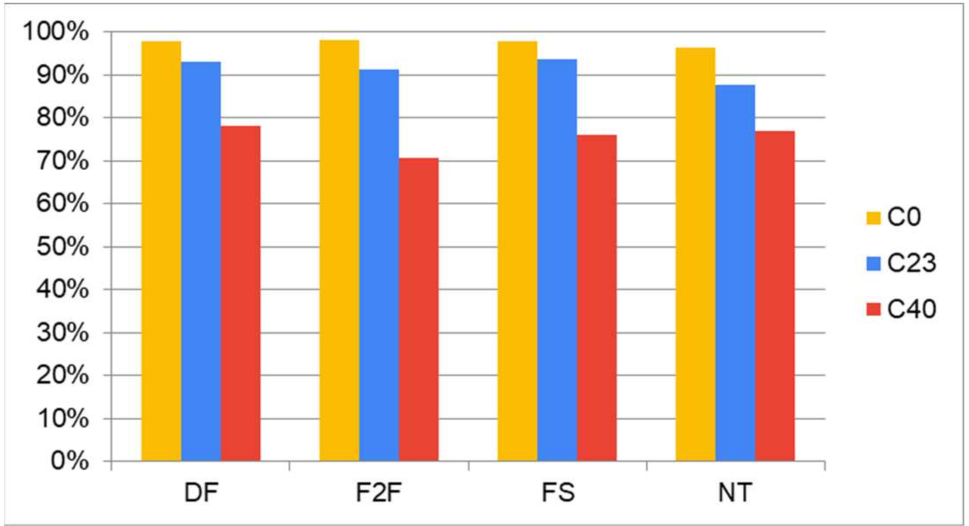




REAL DEEPFAKE

Experimental results

- Results in terms of accuracy have been obtained on the whole test set of *FaceForensics++* by considering different manipulations
- Accuracy **higher than 90%** for *FaceForensics++* dataset (Face2Face, DeepFake, FaceSwap, NT).



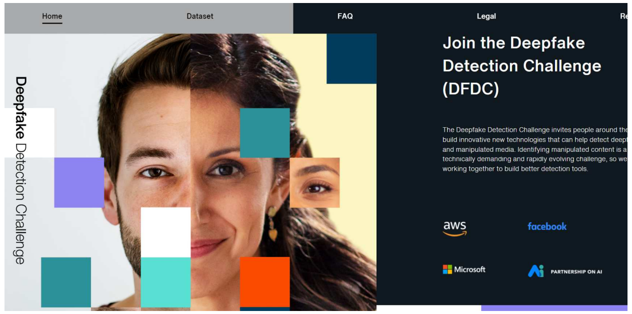
Demo


Score: 1.00
Score: 1.00
THATCHER DIES
Score: 0.00
DeepFakes
TCHER DIES

Future trends

- «Universal method» for forgery detection
  - Independent from kind of manipulations and compressions
  - Deep fake «aware»
  - Multimodal approach is recommended
  - Facebook is investing $ 10M in grants and not only Facebook!!

  https://deepfakedetectionchallenge.ai/
  https://www.kaggle.com/c/deepfake-detection-challenge



- Source identification on Social Media
  - Both device identification and social network provenance need to be examined in depth

Strategies for Countering Fake Information:

new trends in multimedia authenticity verification and source identification

Irene Amerini
amerini@diag.uniroma1.it

Sapienza Università di Roma