AUDIO TAMPERING DETECTION USING MULTIMODAL FEATURES

Simone Milani, Pier Francesco Piazza, Paolo Bestagini, Stefano Tubaro

Dipartimento di Elettronica, Informazione e Bioingegneria, Politecnico di Milano, Milano, Italy

e-mail: {milani@elet./pier.piazza@mail./bestagini@elet./stefano.tubaro@}polimi.it

ABSTRACT

The authenticity verification of a User Generated Audio-Video content relative to a real event can be a very critical task especially when the content is shared on the Internet. Audio-Video files need to be checked in order to verify the origin of the content and the absence of alterations that could have changed their semantic content.

The paper presents a multimodal approach for audio tampering detection that analyzes both the audio component and the video component of a recorded video file. The proposed solution estimates the volumetric characteristics of the environment where the multimedia content has been captured both from the video and audio signals. Then, the approach checks the consistency of the environment characteristics estimated from the audio signal with respect to those estimated from video files. The proposed solution proves to be useful in identifying video fakes and bootlegs, although it proves to be useful for the localization of added audio effects in a movie or radio track.

1. INTRODUCTION

Every day thousands of real events, video messages, artistic performances, newscast or videoblog posts are recorded by a video camera and distributed by uploading the recorded material on data sharing web sites[1]. These contents are generated by different users that autonomously acquire and upload the digital video file by themselves. Since the originating environment and users can not be controlled, it is impossible to state a-priori whether a given video content is a fake (i.e., it has been tempered) or it is original. One of the most frequent alterations that is performed on video files regards the audio track, which can be replaced (e.g., lip dub, studio-edited video, etc.) or altered (e.g., fake sounds, canned laugh and applauses, etc.). Detecting and localizing these changes proves to be extremely useful in both forensic (e.g., bootleg or fake video detection, copyright violations, etc.) and in generic automatic audio editing applications (e.g., canned laugh removal, etc.).

Several forensic analysis strategies have been designed in order to detect alterations on video [2, 3] and audio [4] contents. These rely on revealing traces ("footprints") left on the video signal by different processing steps, and their combination permits an accurate tamper detection. Unfortunately, both video and audio forensic detectors prove to be extremely weak in presence of medium-high compression levels.

The proposed solution aims at adopting a multimodal approach that combines the forensic analysis of the audio component with that of the video component. The final aim of the detector is to distinguish whether the audio track of the analyzed video sequence has been altered with respect to an originally-acquired signal. The detection process relies on an independent estimation of the acquisition environment characteristics from visual and audio data. Whenever inconsistencies can be found between the two estimations, it is possible to conclude that some alteration has been applied since the initial recording. Notice that fusing information from both audio and video data is a technique that proves to be very effective to increase the robustness of a detector [5]. For this reason, when audio and visual information about the recording environment are compatible, it is possible to obtain a more accurate environment estimation than that obtained using a single clue (i.e., only audio or video).

The rest of the paper is organized as follow. Section 2 overviews other works presented in literature on the matter, and Section 3 presents the general scheme of the adopted solution. Section 3.1 describes the environment estimation strategy for video signals, while Section 3.2 presents the environment estimator for audio tracks. Section 4 presents the experimental results obtained for different environments and Section 5 draws the final conclusions.

2. RELATED WORKS

Acquiring, editing, storing, and transmitting an image or video are nowadays extremely-easy tasks due to the widespreading of portable devices (like smart phones and tablets) equipped with a video camera and to the availability of multimedia data editing softwares. These facts have urged the need for effective forensic analysis algorithms that detect alterations and permits validating the authenticity of a given content.

Some of the proposed solution focus on the artifacts left by the acquisition hardware or algorithms. Some of them tries to recover the Photo Response Non Uniformity pattern [6, 7]. Other approaches rely on estimating whether one or more compressions have been applied on the video signals [8, 3]. Other strategies aims at identifying the acquisition device by detecting the type of adopted codec [9] or the specific algorithm that was used [10, 11]. Detecting double compression is also useful whenever frame cuts need to be detected since the statistics of bit rate is altered after recompression [12]. Other solutions rely on the statistics of residual signal that according to recompression changes introducing different artifacts on the reconstructed sequence [13]. Alternatively, other algorithms search for traces left by specific editing operations [2].

Forensic audio strategies have been studied for a long time [14]. The vast majority of these solutions aims at enhancing parts of audio signals, detecting sound sources, and identifying scene or places from audio tracks that sound highly corrupted by noise [15]. More recently, forensic strategies have been concerned with the detection of recompression [4] and fake quality [16]. Other approaches in the literature aims to infer pieces of information related to the size/geometry of the room in which the audio track was recorded.

The project REWIND acknowledges the financial support of the Future and Emerging Technologies (FET) programme within the Seventh Framework Programme for Research of the European Commission, under FET-Open grant number:268478.

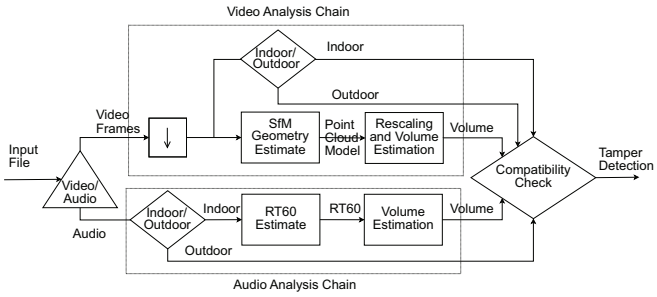

Fig. 1. Block diagram of the proposed scheme

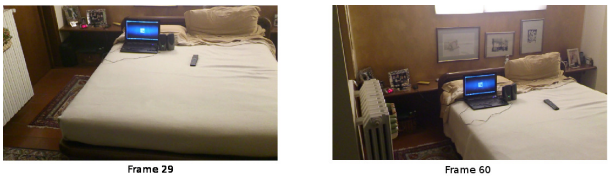

Fig. 2. Frames from dataset F3.

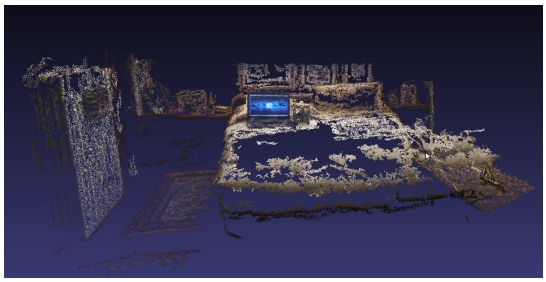

Fig. 3. 3D point cloud model computed from frames of dataset F3.

To this purpose, in [17] a set of audio features is used to characterize specific room types.

However, at the best of authors' knowledge, forensic detectors fusing audio and video information have not been proposed yet. Since multimodal fusion has proven to be a promising methodology [5], in this paper we extend an approach that estimates the characteristics of the environments from reverberation time [18] combining it with the estimate obtained from video information. In doing so, we are able to both have a more accurate estimation of the recording environment and exploit the possible mismatching of audio and visual data to detect the presence of forgeries (i.e., audio track not compatible with the video one).

3. THE PROPOSED APPROACH

The proposed detectors can be divided into two separate analysis chains whose results are combined in the end in order to validate the analyzed video. These two chains are reported in Fig. 1 and process the audio and the video component of the investigated multimedia file separately.

As for the video processing branch, the first unit dump a subset of frames from the video sequence and performs a first classification on them. More precisely, the detector initially discriminates whether the recorded scene has been acquired indoor or outdoor. Then, in case camera is moving around the scene, the initial indoor/outdoor estimate is refined by computing the geometry of the scene via a Structure from Motion (SfM) algorithm applied on subsampled frames. The generated 3D point cloud model is then rescaled according to some known distances that are present in the scene. After this operation, it is possible to estimate the volume of the scene.

Similarly to the video processing chain, the first unit of the audio analysis discriminates whether the scene has been recorded indoor or outdoor. In case the recording environment is classified indoor, it is possible to check whether the signal contains vocalized parts or not. Vocalized parts allow a precise computation of the reverberation time RT60, which permits estimating the volume of the room given an approximated absorption coefficient for the room surface.

Final results are then compared in order to verify that all the different estimates are compliant. In the following we will describe the two analysis chains in detail.

3.1. Environment estimation from video sequence

The frames of the acquired video sequence are sampled in order to reduce the amount of processed data and to avoid analyzing information that is too redundant. The latter requirement proves to be crucial for the SfM geometry estimation unit that needs to process different views of the scene. Assuming that camera is moving, adjacent frames along the time axis could present a reduced amount of innovation, and therefore, they would not improve the geometry estimation. From these premises, we reduce the frame rate of the sequence of $1/10$ in order to keep only those frames with significant motion (see Fig. 2).

It is possible to split the video analysis into three classification units. The first units classify whether the signal has been acquired indoor or outdoor. The second unit computes a 3D point cloud model of the scene whenever the motion of the camera allows it. The third unit map the 3D model to a volume estimate. Further details are reported in the following subsections.

3.1.1. Indoor/outdoor classification

At the beginning of video analysis, the detector needs to decide whether the scene was recorded indoor or outdoor. This analysis permits inferring a minimum level of information about the acquisition environment that could be useful whenever the following geometry estimation fails (i.e., whenever camera is too static or few correspondences can be found between frames). The classification is performed following the strategy reported in [19], i.e., computing a color histogram for the processed frames and classifying it via a Support Vector Machine (SVM) detector.

The input frames are converted into the HSV color space. The H component is quantized into 8 levels, while the S component is quantized into 4 levels. This defines $8 \times 4 = 32$ possible color configurations for each sample in the image. A 32-bins histogram of color configurations is computed and then processed by a binary SVM classifier with Gaussian kernel that has been trained following a RANSAC procedure on a training set of images. The training set of images is a subset of the COREL image database [20].

3.1.2. Geometry estimation via Structure-from-Motion

For a more detailed estimation of the recording environment, the proposed detector resorts to a Structure-from-Motion (SfM) 3D estimation of the scene [21]. More precisely, whenever the camera is

panning around the scene, the different frames can be considered as different views that permit estimating the 3D location of the different objects. To this purposed we adopted the implementation VisualSfM (available at [22]). The output data is a cloud of sparse 3D points that need to be interpolated in order to have a sufficiently-dense set of 3D elements (see Fig. 3). In this set, we need to find some elements related to known quantities in order to scale up the resulting 3D model to the its real dimensions (see the following subsection). As a consequence, data are resampled using the CMVS algorithm [23] available at [24].

3.1.3. Computing the final volume

In the end, data need to be rescaled to the real size. During the estimation process, the SfM algorithm recomputes the focals of the cameras since in the video file this information may be lost. The estimated values may differ from the real one, and as a consequence, the dimensions of objects reprojected in the 3D will result rescaled. To avoid this we identify in the scene some known quantities like the average height of people, the average width of shoulders, the distance between the eyes in a face, and the height of doors.

After the rescaling, the volume of the space is computed by evaluating the maximum and the minimum values of geometrical components for 3D points. From these parameters, it is possible to estimate the box volume that enclose the scene.

3.2. Environment estimation from audio files

In parallel with environmental estimation from video frames, the detector runs an environment detection analyzing the reverberation characteristics of the audio signal. At first, a first discrimination between outdoor and indoor environment is operated. Then, in case the signal has been acquired indoor, a rough estimation of the volume is performed.

3.2.1. Estimation of the reverberation time

An emitted sound diffusing in room presents short or long tails that gradually smooth into silence. These refers to the reflections of the signal on the walls that define the environment and on the objects present in the room.

Traditionally, the late decay envelope has been modeled as an exponential with a single time-constant referred to as decay rate. A common way to characterize the decay rate is to express it as reverberation time, named t_{60} or RT60. The RT60 time measures the time taken for the sound level to drop 60 dB below the level at sound cessation. A frequently-adopted method to estimate RT60 in an experimental set up is the Integrated Impulse Response Method proposed by Schroeder where the recorded sound is integrated starting from different time instants.

Attenuation time is usually computed considering known sound sources (impulsive, Gaussian or spoken). Some blind approaches have been proposed in literature (like that by Loellman *et al.* [25]). Moreover, the adopted reflection model implies that the environment is closed (box model). As a matter of fact, whenever the signal is generic and could have been recorded in an outdoor environment the estimation strategy needs to be modified as follows.

3.2.2. Indoor/outdoor classification via reverberation time values

Applying the estimation strategy suggested by Schroeder to a generic signal does not lead to reliable RT60 since it is not known whether the starting hypothesis are verified. Computing Schroeder RT60 measure on signals acquired outdoor via the routines in the PSYSOUND library, it is possible to notice that the estimated values varies much more during time with respect to acquisitions performed indoor. From these premises, it is possible to compute the histogram $H_{60}(t)$ of the RT60 values for the current audio track under analysis and evaluate

$$\mathcal{T}_{60} = \{\text{RT60 s.t. } H_{60}(\text{RT60}) > \overline{p}\}. \quad (1)$$

The cardinality $N_{RT60} = |\mathcal{T}_{60}|$ permits introducing a measurement of the dispersion which distinguishes indoor acquired tracks from the outdoor tracks. Figure 4 reports the values of N_{RT60} and $E[\text{RT60}]$ for different sequences. It is possible to notice that upper right values characterize outdoor acquisitions. From these results, it is possible to design an SVM classifier that given the feature array $[N_{RT60}, E[\text{RT60}]]$ decides whether the sequence was acquired indoor or outdoor

Table 1. Video detector performance

Acquisition	Real size (m^3)	Real Env.	Est. size (m^3)	Est. Env.
C0	32.40	indoor	30.36	indoor
F0	39.15	indoor	38.48	indoor
F4	29.70	indoor	22.53	indoor
F7	27.00	indoor	16.91	indoor
C2	56.70	indoor	44.87	indoor
F3	71.55	indoor	45.12	indoor
F5	54.00	indoor	79.91	indoor
F6	54.00	indoor	33.06	indoor
F9	54.00	indoor	39.31	indoor
C4	129.60	indoor	71.97	indoor
F1	132.30	indoor	64.17	indoor
OD1	–	outdoor	1078.22	outdoor
OD2	–	outdoor	2727.10	outdoor
OD8	–	outdoor	3702.97	outdoor

3.2.3. Volume estimation

In case the audio track has been acquired indoor, the reverberation conditions are more favorable to an accurate estimation of RT60. Therefore, the RT60 value is refined via the approach in [25] and its value can be related to the volume V of the room via

$$\text{RT60} = \frac{0.161 \cdot V}{S \cdot A_{abs}}, \quad (2)$$

where S is the overall surface, and A_{abs} the absorption coefficient. Experimental results show that, using average absorption coefficients, it is possible to estimate the volume of the room with a certain accuracy for spoken signal. For generic signals, it is only possible to discriminate between small rooms and medium-wide rooms (as it will be shown in Section 4).

4. EXPERIMENTAL RESULTS

The proposed detector works by synergically merging clues from audio and visual data. However, some clues might not be correctly estimated in some cases. An example is the volumetric estimation from video data when the camera is fixed. In this case, the detector must work using only visual information about the indoor/outdoor classification together with audio information. For this reason, we decided to evaluate the accuracy of our detector on different test sets that accommodate different working conditions.

The first test set consists of 23 video sequences recording a known environment (both indoor and outdoor) where a controlled audio signal is emitted. The controlled audio signal presents both

Table 2. Audio performance

Environment	Real Size (m^3)	Real Env.	RT60 (s)	Est. size (m^3)	Est. Env.
C0	32.40	indoor	0.70	154.44	indoor
F0	39.15	indoor	0.50	37.67	indoor
F4	29.70	indoor	0.39	16.75	indoor
F7	27.00	indoor	0.30	7.92	indoor
C2	56.70	indoor	0.568	133.56	indoor
F3	71.55	indoor	0.50	37.67	indoor
F5	54.00	indoor	0.56	57.39	indoor
F6	54.00	indoor	0.49	35.09	indoor
F9	54.00	indoor	0.41	19.53	indoor
C4	129.60	indoor	0.62	87.24	indoor
F1	132.30	indoor	0.69	143.58	indoor

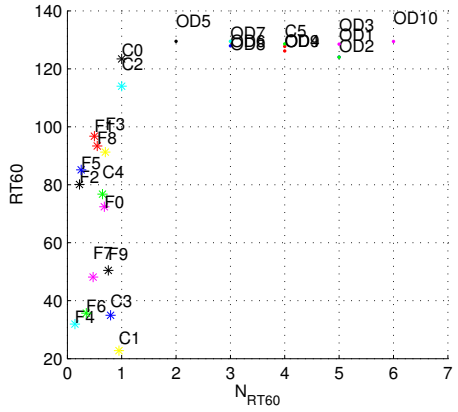

Fig. 4. Indoor/Outdoor detector using features from RT60 estimates. Indoor acquisitions (star) are compared with outdoor acquisitions (dots). Note that the value N_{RT60} for indoor acquisitions is randomly moved between values $[0, 1]$ to make the graph understandable (real values 1).

vocalized and generic (musical) parts. In the environment the camera is panning around the scene making possible to estimate the 3D environment using SfM. Audio sequences usually lasts 40 s, while around 100 video frames are used for the volume estimation. Table 1 reports the estimated volume from the point cloud 3D model on this set. Note that indoor/outdoor classification can be performed by simply thresholding the estimated volume value. Moreover, even though the 3D estimation is not always precise, it is possible to discriminate a rough approximation of room size which can be small, medium, or large (in the indoor scenario).

As it comes to the audio analysis, indoor/outdoor detector works very well since it is possible to obtain an accuracy of 100 % as Fig. 4 shows. Table 2 reports additional details about the estimated volume from RT60 values from indoor acquisition. It is possible to see that the audio detection performs very well whenever the aim of the estimation is to detect a rough approximation of the size of the room. Whenever RT60 is lower than 0.55 s, it is possible to infer that the room is small-medium (with the only mistake of C0). Volume estimation proves to be much more inaccurate than the video case. Anyway, remember that absorption coefficients is not known and therefore, a precise estimation of the volume is not possible.

The second test set involves generic videos downloaded from Youtube, representing both outdoor and indoor video scenes, whose audio track is authentic (i.e., recorded outdoor or indoor according to the scene). In this case, the signal is quite generic and completely uncontrolled by us. Moreover, only in some cases the camera is moving so that SfM algorithm obtains a meaningful 3D estimate of

Table 3. Fake detector performance

Acquisition	Real Env.	Video Est.	Audio Est.	Fake or altered
`ufo_haiti`	outdoor	outdoor	indoor	yes
`catedral_in`	indoor	indoor	indoor	no
`catedral_out`	outdoor	outdoor	outdoor	no
`catedral_vidovic`	indoor	indoor	indoor	no
`catedral_asuncion`	outdoor	outdoor	indoor	yes
`kungfu_bear`	outdoor	outdoor	outdoor	no
`micromax`	outdoor	outdoor	outdoor	no
`nokia`	outdoor	outdoor	outdoor	no
`catedral_bianco`	indoor	indoor	indoor	no

Table 4. Estimated RT60 values (indoor setup)

Env.	Video	Audio est.			
	est.	orig.	M	N	T
room1	0.135	0.129	0.127F	0.126F	0.126F
room2	0.132	0.127	0.127	0.125F	0.125F
room3	0.127	0.123	0.124	0.121F	0.121F
room4	0.129	0.125	0.122F	0.121F	0.121F
room5	0.133	0.128	0.126F	0.125F	0.125F

the scene. As a matter of fact, in most videos the only classification that is possible is to evaluate whether the signal has been acquired indoor or outdoor. Anyway, the proposed SVM based classifier works very well permitting a correct detection in 90 % of the cases. The audio indoor/outdoor classifier works also in this case leading to an accuracy around 95 %. Results are reported in Table 3.

In the end, we considered the possibility of detecting fakes, i.e., sequences where the original audio track acquired on the spot has been replaced with another track that has not been acquired there. It is possible to notice that the designed approach is able to distinguish those videos that have been altered among the set of authentic ones. To this purpose we generated a set of indoor video sequences where the audio tracks have been tampered by adding an environmental noise track. Noise tracks involve background music (M), birds and natural sounds (N), and road traffic (T). Videos are available online and their URLs are reported on [26].

In this case, since videos are indoor and allow volume estimation, in case both audio and video are classified as indoor, it is possible to estimate the conformity of reverberation estimated from the audio track and those approximated from the video signal. Table 4 reports the RT60 values estimated from geometry and from the different audio tracks. It is possible to see that fakes are easily detected verifying if the difference between RT60 values estimated from video and audio is greater than 4 %. Letter F denote sequences that have been detected as fakes. Note that fake detection accuracy is around 90 %.

5. CONCLUSIONS

The paper presented a multimodal approach to detect tampered audio from video sequences. More precisely, the 3D acquisition environment is estimated in parallel from the video data and from the audio data. Experimental results show that the approach is able to estimate whether a sequence has been recorder indoor or outdoor. This capability is also verified on generic sequences downloaded from the web. Future work will be devoted to improve the detector with a fine localization of the tampered part.

6. REFERENCES

[1] S. Milani, M. Fontani, P. Bestagini, M. Barni, A. Piva, M. Tagliasacchi, and S. Tubaro, "An overview on video forensic," *APSIPA Transactions on Signal and Information Processing*, vol. 1, pp. 1–18, 2012.

[2] P. Bestagini, S. Milani, M. Tagliasacchi, and S. Tubaro, "Local tampering detection in video sequences," in *Proceedings of 15th IEEE International Workshop on Multimedia Signal Processing (MMSP 2013)*, Sept 2013, pp. 488–493.

[3] S. Milani, P. Bestagini, M. Tagliasacchi, and S. Tubaro, "Multiple compression detection for video sequences," in *Proceedings of 14th IEEE International Workshop on Multimedia Signal Processing (MMSP 2012)*, Sept 2012, pp. 112–117.

[4] M. Qiao, A. H. Sung, and Q. Liu, "Revealing real quality of double compressed MP3 audio," in *Proceedings of the international conference on Multimedia (MM 2010)*, 2010, pp. 1011–1014.

[5] E. D'Arca, A. Hughes, J. Hopgood, and N. Robertson, "Video tracking through occlusions by fast audio source localisation," in *Proceedings of IEEE International Conference on Image Processing (ICIP 2013)*, Sept 2013, pp. 2660–2664.

[6] W. V. Houten and Z. Geradts, "Source video camera identification for multiply compressed videos originating from youtube," *Digital Investigation*, vol. 6, pp. 48–60, 2009.

[7] K. Rosenfeld and H. T. Sencar, "A study of the robustness of PRNU-based camera identification," in *Proceedings of the SPIE: Media Forensics and Security*, vol. 7254, 2009, pp. 72 540M–72 540M–7.

[8] S. Milani, M. Tagliasacchi, and M. Tubaro, "Discriminating multiple JPEG compression using first digit features," in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2012)*, March 2012, pp. 2253–2256.

[9] P. Bestagini, A. Allam, S. Milani, M. Tagliasacchi, and S. Tubaro, "Video codec identification," in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2012)*, March 2012, pp. 2257–2260.

[10] M. Sorell, "Video provenance by motion vector analysis: A feasibility study," in *Proceedings of International Symposium on Communications Control and Signal Processing (ISCCSP 2012)*, May 2012, pp. 1–4.

[11] S. Milani, M. Tagliasacchi, and S. Tubaro, "Identification of the motion estimation strategy using eigenalgorithms," in *Proceedings of IEEE International Conference on Image Processing (ICIP 2013)*, Sept 2013, pp. 4477–4481.

[12] M. C. Stamm and K. J. R. Liu, "Anti-forensics for frame deletion/addition in MPEG video," in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2011)*, May 2011, pp. 1876–1879.

[13] G. Chetty, M. Biswas, and R. Singh, "Digital video tamper detection based on multimodal fusion of residue features," in *Proceedings of International Conference on Network and System Security (NSS 2010)*, Sept 2010, pp. 606–613.

[14] R. Maher, "Audio forensic examination," *IEEE Signal Processing Magazine*, vol. 26, no. 2, pp. 84–94, March 2009.

[15] B. E. Koenig, "Authentication of forensic audio recordings," *Journal of Audio Engineering Society*, vol. 38, pp. 3–33, 1990.

[16] R. Yang, Y.-Q. Shi, and J. Huang, "Defeating fake-quality MP3," in *Proceedings of the 11th ACM workshop on Multimedia and Security (MM&Sec 2009)*. New York, NY, USA: ACM, 2009, pp. 117–124.

[17] N. Peters, H. Lei, and G. Friedland, "Name that room: room identification using acoustic features in a recording," in *Proceedings of ACM international conference on Multimedia (MM 2012)*, 2012, pp. 841–844.

[18] H. Malik and H. Farid, "Audio forensics from acoustic reverberation," in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2010)*, March 2010, pp. 1710–1713.

[19] O. Chapelle, P. Haffner, and V. N. Vapnik, "Support vector machines for histogram-based image classification," *IEEE Transactions on Neural Networks*, vol. 10, pp. 1055 –1064, 1999.

[20] D. Tao, "The corel database for content based image retrieval," in *https://sites.google.com/site/dctresearch/Home/content-based-image-retrieval*, 2012.

[21] C. Wu, S. Agarwal, B. Curless, , and S. M. Seitz, "Multicore bundle adjustment," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2011)*, June 2011, pp. 3057–3064.

[22] C. Wu, "Visualsfm: A visual structure from motion system," in *http://homes.cs.washington.edu/ ccwu/vsfm/*, 2011.

[23] Y. Furukawa and J. Ponce, "Accurate, dense, and robust multi-view stereopsis," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, pp. 1362 – 1376, 2010.

[24] Y. Furukawa, "Clustering views for multi-view stereo (cmvs)," in *http://http://www.di.ens.fr/cmvs/*, 2011.

[25] H. W. Loellmann, E. Yilmaz, M. Jeub, and P. Vary, "An improved algorithm for blind reverberation time estimation," in *Proceedings of International Workshop on Acoustic Echo and Noise Control (IWAENC 2010)*, Aug 2010.

[26] S. Milani. (2013) Supporting material for Audio Tampering Detection Using Multimodal Features. [Online]. Available: http://www.dei.unipd.it/∼sim1mil/audioTamp