Spotting Audio-Visual Inconsistencies (SAVI) in Manipulated Video

Robert Bolles, J. Brian Burns, Martin Graciarena, Andreas Kathol, Aaron Lawson, Mitchell McLaren
SRI International

Thomas Mensink
University of Amsterdam

Abstract

This paper[1] is part of a larger effort to detect manipulations of video by searching for and combining the evidence of multiple types of inconsistencies between the audio and visual channels. Here, we focus on inconsistencies between the type of scenes detected in the audio and visual modalities (e.g., audio indoor, small room versus visual outdoor, urban), and inconsistencies in speaker identity tracking over a video given audio speaker features and visual face features (e.g., a voice change, but no talking face change). The scene inconsistency task was complicated by mismatches in the categories used in current visual scene and audio scene collections. To deal with this, we employed a novel semantic mapping method. The speaker identity inconsistency process was challenged by the complexity of comparing face tracks and audio speech clusters, requiring a novel method of fusing these two sources. Our progress on both tasks was demonstrated on two collections of tampered videos.

1. Introduction

Videos with audio are becoming a dominant means of documenting events and communicating messages around the world. Modifying or replacing the audio, or replacing the video, is often quite easy to do. These manipulations can change the message drastically, while being, at least in the case of audio, difficult for people to detect. These modifications, however, often leave discrepancies between the visual and audio channels that can be exposed by physical and semantic level analysis.

SAVI (Spotting Audio-Visual Inconsistencies) is a system that we are developing to detect and characterize multiple types of inconsistencies involving different aspects of a video, and fuse these detections into a combined media integrity score. Table 1 lists the inconsistencies being studied, with the types explored in this paper shown in bold. Picture a typical video capturing a person talking or some event happening. It is captured in an environment, such as indoors in a small room, or outdoors on a busy street. The person speaking could be a well-known individual or someone observed earlier in the video. As the person speaks, their speech sounds and lip motions produce distinct patterns associated with what they are saying, and their heads may move relative to the microphone in ways that affect the sound. Finally, in the scene, there could be distinct activities with predictable appearances and sounds, which can either be explicitly classified (e.g., crowds) or produce visual and audio signatures that have strong associations. In each of these aspects of the video, the audio and visual channels must agree, unless the video has been manipulated in some way. The detection of multiple disagreements is a strong indication of some sort of manipulation.

Type of inconsistency	*Visual Features*	*Audio Features*
Environmental class	**Scene: indoor vs. outdoor, small room, etc.**	**Environmental classes, reverberation of closed vs. open spaces**
Speaker identity	**Face recognition**	**Speaker ID**
Lip movement	Pattern of lip movement	Speech patterns
Head movement	Change in head pose	Left-right channel balance
AV device movement	Motion relative to scene	Changes in environmental features
Missing sound	Presence of sound-producing activity	Presence of corresponding sound
Middle-level features	Visual scene signature	Audio scene signature

Table 1. Different types of audiovisual inconsistencies to be detected and characterized in SAVI. The aspects explored in this paper are shown in bold.

Inconsistencies at the levels of environment class and speaker identity produce very different types of manip-

[1]This research was developed with funding from the Defense Advanced Research Projects Agency (DARPA). The views, opinions and/or findings expressed are those of the author and should not be interpreted as representing the official views or policies of the Department of Defense or the U.S. Government.

ulation evidence and distinct challenges detecting them. This paper focuses on the issues associated with these two aspects and the methods we developed to meet them. We also developed two collections of manipulated and non-manipulated videos to evaluate our progress, both of which will be discussed. Our examples were generated by manipulating videos from other, available video collections. The video data cannot be redistributed, but complete instructions for reconstructing our examples from the original collections can be found at http://medifor-av-tampering.ai.sri.com/.

Our original contributions to the field of media forensics include (1) the detection of inconsistencies in audio and visual scenes, (2) the detection of inconsistencies in the changes of audio and visual speaker identity within a video and (3) the design of collections to evaluate these tasks.

2. Related work

Various works have developed methods of audiovisual analysis for the purpose of video indexing and retrieval [6], event detection [10], [22] and video description [24], but not in audiovisual scene inconsistency detection. In work specific to people in video, there has been progress in detecting dubbing via inconsistencies between audio speech patterns and mouth motion [16]. There has also been audiovisual work on determining and localizing the speaking people in a video by means of audio and visual clustering and tracking [15], but none addressing the detection of audiovisual speaker inconsistencies. Components of our system are related to previous work, including visual scene recognition [25],[1], audio speaker recognition [5], the detection and tracking of faces in video [7], face recogntion [20] and facial landmark registration [11].

3. Audiovisual scene inconsistency detection

The goal of the audiovisual scene consistency detection is to verify that the audio characterization of a scene in a video clip taken by a camera is consistent with its visual characterization. For example, if the reverberation and echo properties of the audio track indicate that it was recorded inside a small room, does the visual analysis agree after estimating the distances to the objects in the scene and/or recognizing a small room, such as an office or kitchen? If not, the video may have been manipulated.

3.1. Audio analysis: acoustic scene detection

The acoustic scene detection system was based on i-vectors modeled using a Gaussian backend. The number of Gaussians is based on the number of classes. Mel frequency cepstral coefficients (MFCC) of 20 dimensions were used as audio features extracted from 25ms windows every 10ms. Context was provided with deltas and double deltas resulting in 60 dimensional features. The 400-dimensional i-vector extractor leveraged a 1024 Gaussian Universal Background Model (UBM). I-vectors were projected to 200 dimensions using Linear Discriminant Analysis (LDA). Finally, a Gaussian model was estimated with the i-vectors belonging to each class.

The acoustic scene detection models were trained on a subset of the Placing set collection [4] of about 600 videos, which is a subset of the YFCC100M collection [23]. The training and evaluation sets did not share videos, specific scenes or people.

3.2. Visual analysis: visual scene detection

The visual classification system is based on a zero-example scene detection system, inspired by [12], [1], [13], [19], [17]. The reasoning is twofold, first we observe that there are many more annotated images than videos, and in such a setup we use the images directly to train a highly discriminative deep convolutional net [14, 3, 21]. Second, the scene detection system enables us to align the video and audio annotations semantically, i.e., the classes used in video annotations and audio annotations are different in terms of choice of classes, semantic definition of classes, and dataset collection and annotation procedures. The chosen setup allows us to exploit the existing image annotations and to bridge video and audio annotations.

The core scene recognition system is a frame-based ConvNet model, pre-trained on the recent Places2 dataset [25]. At test time, we sample a frame every 2 seconds and predict the place based on the ConvNet trained on images, see Fig 1 for an example. To obtain a prediction from the full video, the place predictions are averaged over the sampled frames.

To infer the scene, we use a weighted combination of place predictions and place-scene affinity. The affinity between a place and a scene is determined using the Word2Vec [18] distance between the two concepts. Even though the affinity function between places and scene appears simplistic, it seems to encode the affinity well when trained on a large collection of textual data, such as Wikipedia or the meta data of YFC100M.

3.3. Scene inconsistency detection

To develop and test these modules we plan to generate large development and test sets with ground truth annotations. Our top-level plan to do this is to develop Python scripts to cut and paste audio and visual tracks together to form new videos and annotations from annotated source videos. A relatively straightforward example of this would be to extract the audio and visual data from a clip labeled as a *desert* and replace half of the audio with audio from a clip labeled as *indoor-meeting*.

We selected 1,000 videos from the Yahoo Flickr Creative Commons 100 Million (YFCC100m) dataset (http:

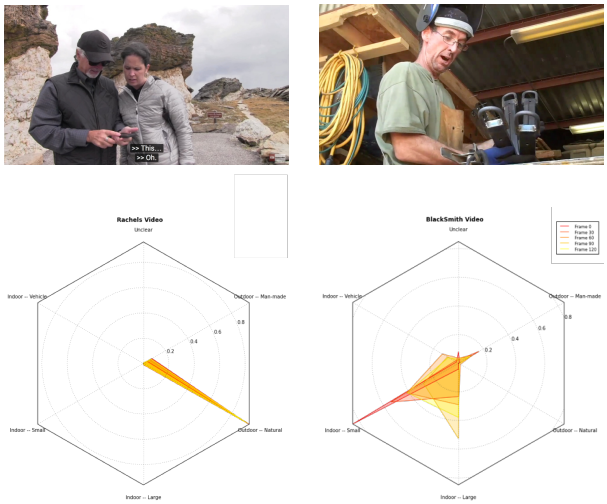

Figure 1. Illustration of visual scene prediction over time.

//yfcc100m.appspot.com/) because the dataset includes a number of annotations, most significantly the scene class and the rank within that class according to the confidence of detection of class membership. This confidence is based on visual features alone, which may have no bearing on the acoustic properties of a given file. Nevertheless, this ranking has been used to guide the selection process.

We developed a set of scripts to automatically construct sets of development and test videos from the YFCC100M data set. Our initial filter checks each video to see if it has an audio track because our techniques require both audio and visual data. Videos without audio were excluded them from further consideration.

An initial spot check revealed that the class label was not always reliable, hence we added a manual review process. Traversing the video files from most to least highly ranked according to (visually determined) class membership, the following annotation was performed:

- list of audio elements in scene
- whether scene video is an instance of class
- whether audio contains elements that will allow detection

Thus, a scene video might rank highly on class membership but if it contains few or no audio elements that would allow classification, it is not suitable for this task. In addition, a judgment is recorded whether a video has been mislabeled, either in terms of its class membership or with respect to major category (e.g., a video set in the outdoor portion of a cafe would misleadingly be grouped into the indoors category).

From the process described above we created two sets:

Untampered video set: this set contains examples from all of the 14 classes that have been judged to contain audio that would make classification possible[2]. One further constraint is that the segment should not contain any diegetic or non-diegetic music (pure soundtrack music already disqualifies a video from consideration since this would mask any scene-internal audio cues).

	False Negatives	False Positives	True Positives	True Negatives	Precision	Recall
Audiovideo tampering evaluation	15	12	35	38	0.745	0.700

Table 2. Results of the audio/video scene tampering evaluation.

Tampered video set: in this initial study of what kinds of scene inconsistencies can be detected, the tampering of scene videos consists of replacing the audio track of a given video with the audio track of a different video from a different category. Future work will include the insertion of audio tracks of different lengths to study the effect of shorter versus longer insertions. For the purposes of tampering, the four categories above have been consolidated into two umbrella categories, "outdoors (outdoor-natural and outdoor-manmade) vs. "indoors (indoor-large and indoor-small). Thus, a tampered video belonging to one of the outdoor classes is paired with the audio of one of the videos of the indoor classes and vice versa. Since videos are not all of the same length, only those video/audio pairings were considered in which the new audio is longer than the original video track so that the audio segment was trimmed to fit the length of the video.

To run the evaluation, we selected the scene class for the visual and audio scene analysis by selecting the most likely class on each models output. Then we compared the class hypothesized by each modality. If they are different then we labeled the video as tampered, otherwise not tampered. The results of the evaluation are detailed in Table 2.

The system correctly identified a number of tampered videos (true positives) and untampered videos (true negatives). For instance, in 2 (a), an outdoor forest scene is paired with the audio track of a room full of people talking and was correctly flagged as tampered. In 2 (b), no tampering is (correctly) detected because the audio characteristics (e.g., reverb) of the audio track appear to match those of the small indoor space detected visually. An illustration of a false alarm is provided in 2 (c), where the audio track can clearly (for humans) be identified as matching the characteristics of the beach scene in the video track, yet the system indicates a mismatch. On the other hand, no mismatch is reported for the example in 2 (d), where

[2]These 14 classes, including 4 categories, are as follows: *beach, forest-path* (outdoor-natural); *market, park, street, train-station* (outdoor-man-made); *cafe, library, living-room, office, restaurant* (indoor-small); *grocery-store, hall, shop* (indoor-large).

Figure 2. Examples of scene inconsistency detection (see text for description of audio): (a) correctly detected tampering, (b) correctly decided as no tampering, (c) falsely detected as tampering, (d),(e) falsely accepted as untampered

the outdoor stream scene is paired with the audio of people speaking indistinctly in a large hall. This last mismatch is similarly not entirely obvious to human observers right away since the flowing water (think of a "babbling brook") is acoustically quite similarly to an undifferentiated set of human voices. In other cases, however, the mismatch can immediately be detected by humans, as for instance in 2 (e) where the indoor video containing speaking people is paired with the audio of birds quietly chirping in a forest. Nevertheless no discrepancy is detected by the system. Clearly the next steps are to increase the training set collection to cover more acoustic conditions as well as improve the modeling techniques and to move towards localizing the inconsistencies along the temporal axis.

4. Audiovisual speaker inconsistency detection

Some manipulations involve replacing a person's speech with someone else's, such as in dubbing, or perhaps replacing the face, but not the audio. These manipulations can produce observable inconsistencies between the identity of the audio speaker and that of the visibly talking face. If we have a database of known subjects with samples of their speech and faces, we can then check for inconsistent audiovisual matches to the database. It is also possible to check for inconsistencies without knowing the actual identity of the people in the video: did the voice of this talking face change during the video, but not the face itself, or vice versa? It is the detection of inconsistent change within a video that is explored in this paper. Our methods can be naturally extended to work with annotated, external sources and also with collections of non-annotated video.

The key to detection here is to find at least two segments (time intervals) for which the audio *or* visual identity of the speaker differ, but not both. We achieved this by processing each modality independently first and then checking for inconsistencies between them. Within each modality, we detected speech segments and estimated the probability that the speaker was the same in each pair of these segments. The probability of inconsistency was then computed for overlapping pairs of audio and visual segments. In this process, we were assuming that there is no crosstalk, in other words, there was only one speaker at a time. In practice, crosstalk can occur and will require us to detect situations where it can interfere with our interpretation.

4.1. Audio analysis: speaker diarization

Speaker diarization is the process of partitioning the audio into segments that are homogeneous with respect to speaker and organizing the segments into clusters with the same speaker characteristics (and hence likely to be the same speaker). The diarization here was based on clustered i-vectors using a PLDA-based i-vector speaker recognition system [5] followed by viterbi realignment. The speaker recognition system was trained from datasets used in the 2004-2008 NIST Speaker Recognition Evaluations. Mel frequency cepstral coefficients (MFCC) of 20 dimensions were used as audio features extracted from 25ms windows every 10ms. Context was provided with deltas and double deltas resulting in 60 dimensional features. The 400-dimensional i-vector extractor leveraged a 1024 Gaussian Universal Background Model (UBM). I-vectors were projected to 200 dimensions using Linear Discriminant Analysis (LDA) followed by mean and length-normalization prior to use in PLDA scoring.

Speaker diarization first involved segmenting the audio into 2 second blocks with 1 second overlap. The speaker recognition system was then used to generate a matrix of scores for the i-vectors from the audio being diarized. These scores are then transformed into a distance matrix. The distances are computed as the opposite of the log-likelihood ratios (LLR) obtained with PLDA shifted by the maximum LLR obtained for any pair of samples. This way, the minimum distance is 0. Finally, hierarchical clustering with average linkage method is used to generate a clustering tree. The tree is then pruned by ensuring that each cluster has a cophenetic distance no greater than a certain threshold t. A value of t=-9.0 optimized clustering performance on the development set and was used to cluster the evaluation data.

The resulting segment-to-speaker distance matrix and the locations of the segments in the video were used to search for inconsistencies with respect to the visual processing output. Using logistic regression, each distance

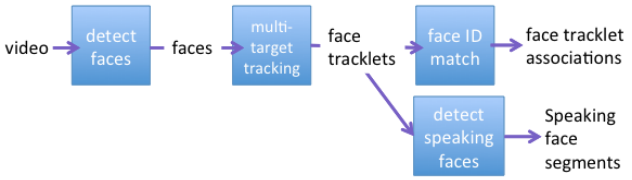

Figure 3. Processing steps for generating speaking face segments and computing their associations (probability of being the same person).

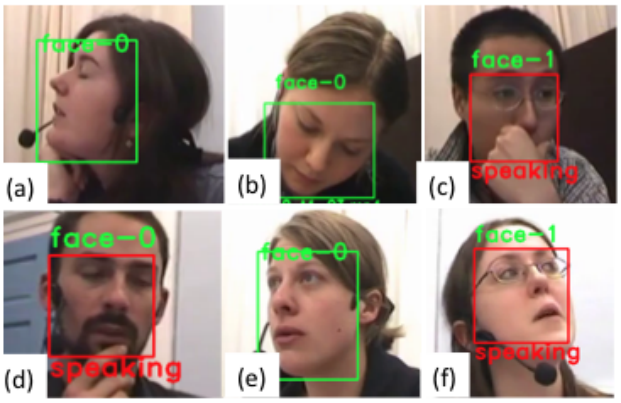

Figure 4. Examples of faces from the AMI meetng collection showing range of poses, occlusions and the similarity of some faces (e.g., b and e). Face labels shown are internal to the separate video examples (no meaning here). All speakers correctly detected except for (a).

was converted to a segment-speaker association probability, $p_{\text{sassoc}}(a, s)$, for audio segment a and speaker s.

4.2. Visual analysis: detecting and tracking talking faces

To analyze the audiovisual consistency of the speakers, we needed to find segments in the video where people were seen talking and determine which talking face segments corresponded to the same person (perhaps seen much later in the video). To accomplish this, our system performed the steps in Figure 3: detect faces in the video, track them, determine which face tracks are of the same person (associate the tracks) and determine when a face is talking or not. Figure 4 shows some examples of faces from the AMI Meeting collection [2] used here to study audiovisual speaker consistency. The meeting data contains many challenging aspects, including a wide range of face poses (Figure 4 a-f), head motions, similar faces (Figure 4 b, e) and occlusions (Figure 4 c, d) (e.g., hands, laptops, microphones).

For face detection, we trained a CNN that was a variant of AlexNet [14] using approximately 1.5 million IMBD faces and applied it in a fully convolutional mode over

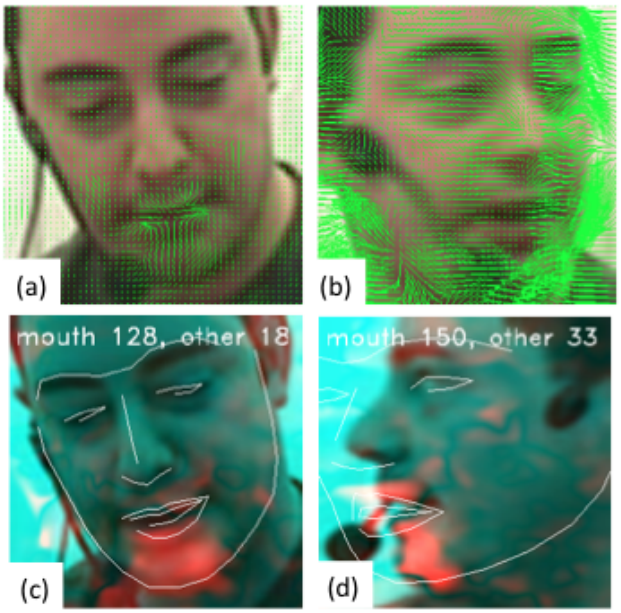

Figure 5. Speech detection using flow: (a) optic flow during speaking, (b) during head motion, (c) aligned face landmarks (white) and vertical flow magnitude (red is high flow, numbers show magnitude in mouth area vs. the rest), (d) example with face in profile and poorly aligned landmarks.

multiple scales with a scale step of two to the quarter power (quarter octave steps). Our tracking process had two parts, similar to the well-demonstrated tracklet-based framework [9]: first find (possibly incomplete) sequences of detections for which the frame-to-frame links have a high probability of being correct (a *tracklet*), then form an association matrix over the tracklets using a similarity measure. For our purposes here, this association matrix did not need be resolved into distinct, extended tracks, but could be used directly by the audiovisual analysis described in the next section. The tracklets were sequences where the adjacent detections were very close in both the image space and in a facial identity feature space (face ID features). The face ID features were computed from a face image using a CNN of GoogleNet design [21]. This was trained on two million face images of approximately sixteen thousand individuals using a combination of softmax and triplet-loss-based training [20]. Using logistic regression, the association probability $p_{\text{fsame}}(f_1, f_2)$ between each pair of face tracklets f_1 and f_2 (the probability of being the same person) was computed from the tracklet face ID features.

Visual speech was detected using a combination of aligned facial landmarks [11] and optical flow (Opencv implementation of Farneback [8]). Figure 5 shows examples of flow and aligned landmarks for AMI faces. Our initial tests on AMI data showed that detecting speech using

the changes in aligned face landmarks, such as the upper and lower lips, does not always work well, especially in face poses that are difficult for current alignment methods, such as profile views (Figure 5 d). This is quite common in videos of interviews and meetings. What worked better during testing on our development set was to use the aligned landmarks to localize the general area of speech production (the mouth and jaw) and measure the vertical flow magnitude in this area relative to other nearby areas, after subtracting out the average flow of the whole face. This method was used in the evaluation. Figure 5 (c-d) shows this flow magnitude where red is high, and Figure 5 (d) shows that the speech area can still be roughly localized even after poor alignment of the details. The detected speech was then averaged over segments in time, and segments with high levels were flagged as talking faces. We found that this approach works well for the meeting data studied, though sometimes detection failed when speech production was harder to observe, such as in Figure 4 (a), where the speaker is resting her chin on her fist and is in profile. For the study reported here, each segment was defined to be the duration of a face tracklet, bounded in time by shot changes or obscurations of the face, and the experiments (discussed below) were designed such that identity changed (if at all) at shot changes. This allowed us to focus the current study on audiovisual inconsistency detection, future work will include more fine-grained visual speech segmentation and a study the temporal limits of inconsistency analysis.

4.3. Speaker inconsistency detection

Figure 6 shows talking face and audio speech segments for different example videos from our evaluation set. Each green row is a talking face segment (distinct face track), and each red row is an audio speaker (speech cluster). Time is horizontal and the tick marks in the middle are seconds. The speaker rows show the speech segments as blocks of red of different brightness, where brightness indicates $p_{\text{sassoc}}(a, s)$, the estimated probability that audio segment a is from speaker s (not normalized across speakers since we do not know if the speakers are actually separate people). Figure 6 (a) and (b) show two examples of inconsistencies: two talking faces that both coincide with the same audio speaker, and vice versa for (b). Figure 6 (c) and (d) are examples without inconsistencies: (c) has two talking faces that each coincide with a distinct speaker, and, even though (d) shows two speaker rows, it is in fact a single talking face and speaking person that happens to be associated with two speech clusters. The occurrence of multiple clusters per actual speaker can be predicted by the degree of overlap in the segments shared by the clusters, as shown in Figure 6 (d). Given this, the following steps were used to estimate the probability of inconsistencies in an example:

1. For every pair of audio speakers (clusters) s_1 and s_2, estimate the probability that they are actually the same person:

$$p_{\text{ssame}}(s_1, s_2) = \sum_a \min(\hat{p}_{\text{sassoc}}(a, s_1), \hat{p}_{\text{sassoc}}(a, s_2)), \tag{1}$$

where,

$$\hat{p}_{\text{sassoc}}(a, s)) = p_{\text{sassoc}}(a, s) / \sum_{a'} p_{\text{sassoc}}(a', s). \tag{2}$$

2. Estimate the probability $p_{\text{coin}}(f, s)$ that talking face segment f and a speaker s coincide in the video by summing $p_{\text{sassoc}}(a, s)$ over all segments a that intersect with f and mapping this to a probability via logistic regression.

3. Given every pair of possibly coinciding faces and audio speakers $[(f_1, s_1), (f_2, s_2)]$, including pairs where the face or the speaker are the same (*i.e.* $[(f, s_1), (f, s_2)]$ and $[(f_1, s), (f_2, s)]$), compute the probability of inconsistency as:

$$\begin{aligned} &p_{\text{inconsis}}(f_1, s_1, f_2, s_2) \\ &\quad = p_{\text{coin}}(f_1, s_1) * p_{\text{coin}}(f_2, s_2) * \\ &\quad (p_{\text{fsame}}(f_1, f_2) * (1 - p_{\text{ssame}}(s_1, s_2)) \\ &\quad + p_{\text{ssame}}(s_1, s_2) * (1 - p_{\text{fsame}}(f_1, f_2))) \end{aligned} \tag{3}$$

The probability of inconsistency for the whole video is taken as the maximum over all pairs. The values shown

Figure 6. Talking face and audio speech segments from four example videos along with computed inconsistency scores: (a) and (b) are inconsistent, (c) and (d) are not (see text for explanation).

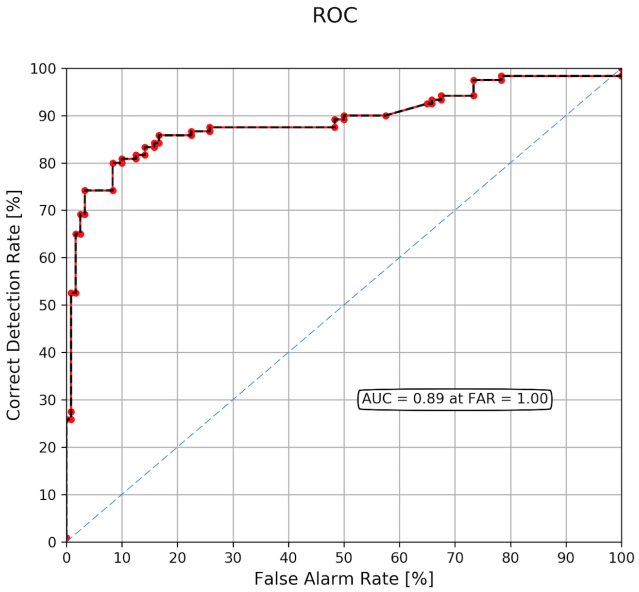

Figure 7. The ROC curve for audiovisual speaker inconsistency detection.

in Figure 6 are the probabilities of inconsistency calculated for each example using this approach.

4.4. Speaker inconsistency experiments

Examples of tampered and non-tampered videos were generated using the AMI Meeting collection [2], which contains recordings of meetings where each participant has a headset microphone and a video camera trained on them. Using ffmpeg we de-interlaced and combined the videos in various ways. For video, we combined the feeds from two cameras, so that each frame has two people. This makes the speaker detection and face-to-face association across shot changes more challenging. Then, we generate untampered videos from either one shot or two shots, but always making sure the speakers in the audio and video coincide (the different shots may have different speakers, or the same speakers, but the modalities are consistent). We generate tampered video by doing something similar, but in this case, the speakers change at some point in the video in one modality but not in the other, generating an inconsistency internal to the video. For our evaluation, we generated a total of 240 examples, half tampered, varying all of the above aspects and varying the subjects involved. Video lengths ranged from 6 to 40 seconds.

Figure 7 shows the resulting ROC curve and computed area under the ROC for the evaluation set. Clearly the system got a number of the videos correctly, but there is room for improvement in detecting talking in faces and estimating the probability of associations and co-occurrences needed to make the decision.

5. Conclusions

This paper presents two methods of detecting potential video tampering by exploiting two types of audiovisual inconsistencies: scene type and speaker identity. Novel aspects of our approach include the method of semantic mapping between mismatched audio and visual scene collections, and the probabilistic audiovisual inconsistency detection from face tracks and audio speaker clusters. Experiments on a new tampered-video collection showed promise for these methods. Future work will improve scene inconsistency detection by increasing the training set to include more acoustic conditions, improving the modeling techniques and doing more fine-grained temporal localization of inconsistencies. More fine-grained localization of speaker inconsistency detection will also be explored, especially in the process of visual speech detection. Speaker inconsistency detection will also benefit from more careful modeling of the probabilities involved. Finally, the two methods presented here are part of a set of tools that will tackle multiple aspects of audiovisual inconsistency, the fusion of which is our central approach to robust tampering detection.

References

[1] S. Cappallo, T. Mensink, and C. Snoek. Video stream retrieval of unseen queries using semantic memory. In *British Machine Vision Conference (BMVC)*, 2016. 2

[2] J. Carletta. Announcing the ami meeting corpus. *The ELRA Newsletter*, 11(1):3–5, 2006. 5, 7

[3] K. Chatfield, K. Simonyan, A. Vedaldi, and A. Zisserman. Return of the devil in the details: Delving deep into convolutional nets. In *BMVC*, 2014. 2

[4] J. Choi, B. Thomee, G. Friedland, L. Cao, K. Ni, D. Borth, B. Elizalde, L. Gottlieb, C. Carrano, R. Pearce, and D. Poland. The placing task: A large-scale geo-estimation challenge for social-media videos and images. In *Proceedings of the 3rd ACM Multimedia Workshop on Geotagging and Its Applications in Multimedia*, 2014. 2

[5] N. Dehak, P. J. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet. Front-end factor analysis for speaker verification. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(4):788–798, 2011. 2, 4

[6] M. Douze, J. Revaud, J. Verbeek, H. Jégou, and C. Schmid. Circulant temporal encoding for video retrieval and temporal alignment. *International Journal of Computer Vision*, 119(3):291–306, 2016. 2

[7] M. Du and R. Chellappa. Face association for videos using conditional random fields and max-margin markov networks. *IEEE Trans. Pattern Anal. Mach. Intell.*, 38(9):1762–1773, 2016. 2

[8] G. Farnebäck. Two-frame motion estimation based on polynomial expansion. In *Proceedings of the 13th Scandinavian Conference on Image Analysis*, SCIA'03, pages 363–370, Berlin, Heidelberg, 2003. Springer-Verlag. 5

[9] W. Ge and R. Collins. Multi-target data association by tracklets with unsupervised parameter estimation. In *Proceedings of the British Machine Vision Conference*, pages 93.1–93.10. BMVA Press, 2008. doi:10.5244/C.22.93. 5

[10] A. Habibian, T. Mensink, and C. Snoek. Video2vec embeddings recognize events when examples are scarce. *Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 2016. 2

[11] P. Huber, G. Hu, R. Tena, P. Mortazavian, P. Koppen, W. Christmas, M. Ratsch, and J. Kittler. A multiresolution 3d morphable face model and fitting framework. In *Proceedings of the 11th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications*, 2016. 2, 5

[12] M. Jain, J. C. van Gemert, T. Mensink, and C. G. Snoek. Objects2action: Classifying and localizing actions without any video example. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4588–4596, 2015. 2

[13] S. Kordumova, T. Mensink, and C. Snoek. Pooling objects for recognizing scenes without examples. In *ACM International Conference on Multimedia Retrieval (ICMR)*, 2016. 2

[14] A. Krizhevsky, I. Sutskever, and G. Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, 2012. 2, 5

[15] N. Le, S. Meignier, and J.-M. Odobez. Eumssi team at the mediaeval person discovery challenge 2016. In *MediaEval Benchmarking Initiative for Multimedia Evaluation*, Oct. 2016. 2

[16] N. Le and J.-M. Odobez. Learning multimodal temporal representation for dubbing detection in broadcast media. In *ACM Multimedia*. ACM, Oct. 2016. 2

[17] T. Mensink, E. Gavves, and C. G. Snoek. Costa: Co-occurrence statistics for zero-shot classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2441–2448, 2014. 2

[18] T. Mikolov, K. Chen, G. Corrado, and J. Dean. Efficient estimation of word representations in vector space. *CoRR*, abs/1301.3781, 2013. 2

[19] M. Norouzi, T. Mikolov, S. Bengio, Y. Singer, J. Shlens, A. Frome, G. Corrado, and J. Dean. Zero-shot learning by convex combination of semantic embeddings. In *International Conference on Learning Representations (ICLR)*, 2014. 2

[20] O. M. Parkhi, A. Vedaldi, and A. Zisserman. Deep face recognition. In *British Machine Vision Conference*, 2015. 2, 5

[21] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. In *Computer Vision and Pattern Recognition (CVPR)*, 2015. 2

[22] N. Takahashi, M. Gygli, and L. V. Gool. Aenet: Learning deep audio features for video analysis. *CoRR*, abs/1701.00599, 2017. 2

[23] B. Thomee, D. Shamma, G. Friedland, B. Elizalde, K. Ni, D. Poland, D. Borth, and L.-J. Li. The new data and new challenges in multimedia research. *Communications of the ACM*, 59(2):64–73, 2016. 2

[24] J. Xu, T. Mei, T. Yao, and Y. Rui. Msr-vtt: A large video description dataset for bridging video and language. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 2

[25] B. Zhou, A. Khosla, À. Lapedriza, A. Torralba, and A. Oliva. Places: An image database for deep scene understanding. Technical report, ArXiV, 2016. 2