International Journal of Technology 10(7): 1335-1343
ISSN 2086-9614 © IJTech 2019

ACOUSTIC PORNOGRAPHY RECOGNITION USING FUSED PITCH AND MEL-FREQUENCY CEPSTRUM COEFFICIENTS

Rasoul Banaeeyan[1*], Hezerul Abdul Karim[1], Haris Lye[1], Mohamad Faizal Ahmad Fauzi[1], Sarina Mansor[1], John See[2]

[1]*Faculty of Engineering, Multimedia University, Cyberjaya, 63100, Malaysia*
[2]*Faculty of Computing & Informatics, Multimedia University, Cyberjaya, 63100, Malaysia*

(Received: November 2018 / Revised: January 2018 / Accepted: September 2019)

ABSTRACT

The main objective of this paper is pornography recognition using audio features. Unlike most of the previous attempts, which have concentrated on the visual content of pornography images or videos, we propose to take advantage of sounds. Using sounds is particularly important in cases in which the visual features are not adequately informative of the contents (e.g., cluttered scenes, dark scenes, scenes with a covered body). To this end, our hypothesis is grounded in the assumption that scenes with pornographic content encompass audios with features specific to those scenes; these sounds can be in the form of speech or voice. More specifically, we propose to extract two types of features, (I) pitch and (II) mel-frequency cepstrum coefficients (MFCC), in order to train five different variations of the k-nearest neighbor (KNN) supervised classification models based on the fusion of these features. Later, the correctness of our hypothesis was investigated by conducting a set of evaluations based on a porno-sound dataset created based on an existing pornography video dataset. The experimental results confirm the feasibility of the proposed acoustic-driven approach by demonstrating an accuracy of 88.40%, an F-score of 85.20%, and an area under the curve (AUC) of 95% in the task of pornography recognition.

Keywords: Acoustic recognition; KNN classifier; MFCC features; Pornography detection

1. INTRODUCTION

Filtering inappropriate visual content from different sources (internet television (TV), web pages, etc.) is a primary concern in environments, such as schools, homes, and workplaces. In some countries, such as Malaysia, Indonesia, and Brunei, all TV channel providers are expected to obtain suitability approval before granting access to their subscribers or public users.

One part of the suitability assessment involves pornography recognition, which, most of the time, imposes a huge censorship cost to the service providers due to the need to recruit a large amount of manpower to work constantly over several months.

The main purpose of this research is to facilitate the task of pornography detection by proposing to exploit the distinctive power of acoustic features (as explained in Section III). More specifically, this study proposes employing pitch and mel-frequency cepstrum coefficient (MFCC) acoustic-related features, which represent both voiced and unvoiced sounds.

Although there have been several attempts to address the problem of pornography recognition

*Corresponding author's email: banaeeyan@gmail.com, Tel. +60-17-3490161, Fax. +60-03-8756545
Permalink/DOI: https://dx.doi.org/10.14716/ijtech.v10i7.3270

(Caetano et al., 2016; Geng et al., 2016; Moreira et al., 2016; Nian, et al., 2016; Zhou et al., 2016; Jin et al., 2018; More et al., 2018; Nurhadiyatna et al., 2018; Shen et al., 2018;), almost all of them have utilized visual content to automate the target task of sensitive content detection.

The paper is organized as follows. The next section (2) briefly overviews recent similar works in the domain, followed by Section 3, which presents the design framework of the proposed acoustic-driven pornography recognition, as well as the details of the system design employed in this study. Section 4 details the experimental setup and procedures followed in our research to facilitate the reproducibility of the results. In Section 5, the results of the different experiments are presented and discussed; this is followed by Section 6, which concludes the paper and states some possible future directions.

2. RELATED WORK

The importance of fusing different modalities (video, audio, text, etc.) in the context of multimedia information processing has been confirmed in the literature (Snoek et al., 2005; Snoek & Worring, 2007; Jiang et al., 2013). Yet, the majority of the related works on pornography detection rely on the visual content by extracting frame descriptors using some widely utilized feature descriptors, such as local binary patterns (LBP) (Zhou et al., 2012), scale invariant feature transform (SIFT) (Lowe, 1999), or histogram of oriented gradients (HOG) (Dalal et al., 2005) to obtain local or global video descriptors and later use them to differentiate between pornography and normal content in videos. Several instances of such works include Caetano et al. (2016), Nian et al. (2016), Zhou et al. (2016), Jin et al. (2018), More et al. (2018), Nurhadiyatna et al. (2018), and Shen et al. (2018).

There have been several scholarly attempts in the literature to incorporate acoustic information in order to improve different visual recognition tasks. For instance, Pieropan et al. (2014) proposed enhancing the classification of human actions in videos by using MFCCs of the audio of those videos. To this end, the authors extracted a set of 13 coefficients and used them as inputs to a hidden Markov model (HMM) to train a classifier.

The same MFCC feature (12 coefficients) was used to solve the fifty-class classification problem of different environmental sounds (Piczak, 2015). Piczak (2015) created a new dataset of sounds belonging to main categories, such as animal, nature, and human, and, more specifically, proposed to train three different supervised models as k-nearest neighbor (KNN), Random Forest, and support vector machine (SVM) based on the MFCC coefficients.

Foggia et al. (2015) developed a system to detect audio events from closed-circuit television (CCTV) cameras. A new dataset comprising different short and long event sounds was created, and they proposed extracting three categories of audio features from the dataset: (I) spectral features (spectral centroid, spread, roll-off, and flux), (II) energy features (energy, volume, and sub-band energy ratios), and (III) temporal features (zero crossing rate (ZCR)). Later, the authors trained different SVM classifiers to test the performance of the proposed approach.

MFCC features were also utilized in the task of urban sound classification in a study by Salamon and Bello (2015), where the authors proposed extracting 25 coefficients and used them to train an unsupervised spherical K-means algorithm to solve a ten-class problem. The same features were employed by Stowell et al. (2015) to predict the labels of various scenes from their associated audios by training a continuous density HMM.

To enhance the performance of road accident detection in hazardous situations, Foggia et al. (2016) proposed using MFCC features as discriminative indicators of such incidents. To this end, they extracted features at two different levels: at the low-level, their system extracted distinctive descriptors to capture properties of targeted events, and, at the high-level, a Bag of Words (BoW)

was employed to identify short and sustained events.

In addition, Mesaros et al. (2016) constructed a dataset of environmental soundtracks and trained a supervised classifier by using a Gaussian mixture model (GMM) and MFCC as feature descriptors. Finally, Gemmeke et al. (2017) introduced a large-scale audio dataset of seven main categories, which was created by collecting thousands of ten-second YouTube videos. It was proposed to predict the corresponding labels of each sub-category by training a fully-connected neural network (NN).

There are two relevant studies in the literature that proposed to exploit the distinctiveness power of acoustic information to improve performance in tasks, such as the recognition of blue movies (Zuo et al., 2008) and the identification of objectionable soundtracks. Zuo et al. (2008) presented a model to recognize porno-sounds by training a GMM, while the training samples were the thirteen-dimensional MFCC features of the audios taken from videos.

Shi et al. (2013) proposed identifying objectionable segments of the audio files by employing two newly presented variants of GMM: taking advantage of a novel distance measurement and building mixture models. The authors also conducted different experiments with other supervised learning models, such as pseudo-GMM, heterogeneous mixture models (hetMM), SVM, and artificial neural networks (ANN).

3. METHOD

The overall system design of the proposed acoustic pornography recognition system is depicted in Figure 1. After the soundtracks are extracted from the video dataset, they are categorized into two groups of positive (pornography) and negative (normal) sounds. In the next phase, two different sets of features are extracted from these samples, namely, pitch-based features and MFCC-based features in order to be concatenated and used as inputs to different variants of KNN classifiers.

After the KNN models are trained, at testing time, a sample sound goes through the same process of feature extraction, and, finally, its features are used by the trained model to predict the probability values corresponding to the two classes of pornography and normal sounds; the class with the highest probability is selected as the predicted label.

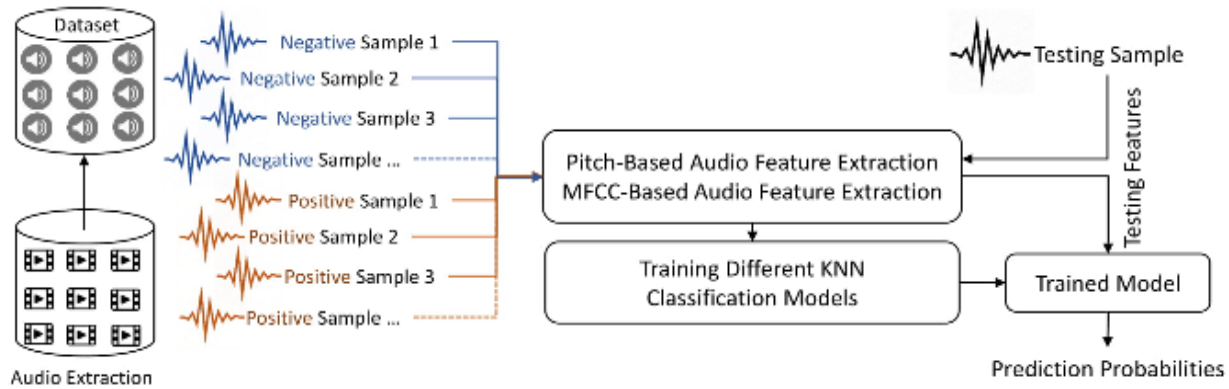

Figure 1 The overall system design of the proposed acoustic pornography recognition system

After the audio files are extracted from the video files, they are divided into two sets of positive (pornography sounds) and negative (normal sounds) samples. Later, two feature extraction algorithms are applied to the samples, and the resultant audio descriptors are concatenated to form a unique representation of the samples. The sampled descriptors, along with their corresponding labels, are fed as inputs to different supervised KNN algorithms to train different classification models.

3.1. Pitch-based Features

Speech is basically classified as voiced and unvoiced. In the former case, vocal cords modulate the air streams in the lungs, and this generates a quasi-periodic excitation. The produced sounds are dominated by a proportionately low frequency oscillation, which is defined as pitch.

It also refers to the degree of lowness or highness of a tone, which is comprehended by human ears. The quantity of vibrations produced by vocal cords (per second) generates various pitches. In this research, we used pitch-based features because they are the acoustic signals that are correlated with intonations and tones in human speech and thus can best present the sounds in pornography videos.

In this paper, pitch features are estimated according to the fundamental frequencies of an input signal at various locations, as determined by: (1) window length; and (2) overlap length. More specifically, per Atal (1972), the normalized correlation function (NCF) is applied to estimate the pitch. The algorithm segments the input signal based on the window length and overlap length (Figure 2) and uses these segments as input arguments; as a result, it outputs the pitch features.

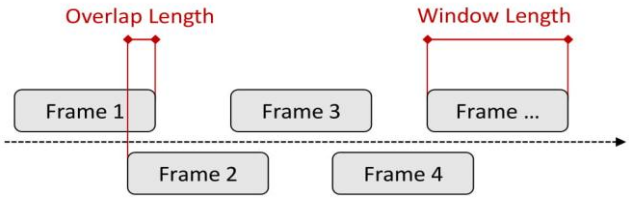

Figure 2 An example presenting the selection of four segments/frames from an input signal with fixed defined window length and overlap length

3.2. MFCC-based Features

MFCCs are features of speech signals, which can be used in the classification of speeches (Hasan et al., 2004). They are capable of representing vocal tracts (noise-like excitations). While the frequency representations of vocal tracts are comparably smooth, those of voiced speech are shaped as an impulse train. The vocal tracts can be modeled by the spectral envelopes in a segment of speech.

In this research, we used MFCC-based features to represent the noise-like excitations in the soundtracks of the pornography videos. The MFCCs enable compressing information related to the vocal tracts to a pre-determined compact number of coefficients.

The MFCC segments the input signal into various overlapped frames and then computes the cepstral features for each individual frame. More specifically, the algorithm outputs the delta, the difference between the previous coefficient and the current coefficient over the defined frame length. The algorithm, as in the study by Rabiner and Schafer (2011), calculated the cepstral values by fitting the coefficients of the adjacent frames of an input signal.

3.3. Training Hyper Parameters

With respect to the pitch feature extraction, the two parameters, window length and overlap length, were respectively set to $Hz \times 52 \times 10^{-3}$ and $Hz \times 42 \times 10^{-3}$, where Hz is the frequency of the input signal. Regarding the MFCC feature extraction, the frame length and overlap length were respectively set to $Hz \times 3 \times 10^{-2}$ and $Hz \times 2 \times 10^{-2}$, and the number of extracted coefficients totaled 15.

4. EXPERIMENTS

This section details the experimental procedure, the choice of the datasets, and the performance metrics, as well as the system specifications.

4.1. Video and Sound Dataset

Although the literature contains instances of audio datasets (e.g., for human activity recognition) (Piczak, 2015), there is still a lack of pornography acoustic dataset. In this research, we used a video dataset presented by Lopes et al. (2009), which included 179 samples of videos; sample video frames are depicted in Figure 3. Out of 179 clips, 169 soundtracks were extracted using a third-party software, while 10 clips were not included in the sound samples due to the corruption of the videos. Overall, the newly created pornography soundtracks included 89 instances representing positive samples and 80 instances representing negative samples.

Figure 3 Presentation of sample pornography and normal content. The instances are taken from the video dataset. The top row presents negative samples, and the bottom row presents positive samples

4.2. Training and Testing Partitions

To train different versions of the KNN models, the audio dataset was randomly divided into two parts for training and testing the performance, with each partition containing 75% (127 audio tracks) and 25% (42 audio tracks) of the images in the collection.

4.3. Performance Metrics

Following previous works in the field (Naik & Metkewar, 2015; Gupta et al., 2016), three different evaluation metrics were used to assess the performance of the proposed binary classification models (nudity vs. non-nudity), namely, *Accuracy*, *AUC*, *and F-score*, as formulated in Equations 1, 2, 3, and 4.

In addition, an receiver operating characteristic (ROC) performance curve was also generated to provide more insight into the behavior of the classifier with respect to different prediction threshold values (ranging from 0 to 1). AUC is the area under the ROC curve, indicating the performance of a binary classifier.

$$Accuracy = (TP+TN)/(TP+TN+FP+FN) \quad (1)$$

$$F\text{-}score = 2\times(Precision\times Recall)/(Precision+Recall) \quad (2)$$

$$Precision = TP/(TP+FP) \quad (3)$$

$$Recall = TP/(TP+FN) \quad (4)$$

where TP is the number of true positive samples (correctly classified as pornography), TN is the number of true negatives (correctly classified as normal), FP is the number of false positives (incorrectly classified as pornography), and FN is the number of false negatives (incorrectly classified as normal).

4.4. Software and Hardware

All of the implementations were carried out using Matlab R2018b (academic version), Image Processing Toolbox, Computer Vision Toolbox, and Statistics and Machine Learning Toolbox. The experiments were conducted on a desktop computer with Windows 7 (64-bit), 16 GB RAM, Intel Core i7-4790 CPU @ 3.60 GHz, and an NVIDIA GPU with 4 GB internal memory (GTX 749).

5. RESULTS AND DISCUSSION

This section presents the experimental results on the newly constructed porno-sound dataset. Table 1 presents the performances of the six different KNN classifiers used in our study with respect to five performance metrics.

Table 1 Performance comparison of different variations of KNN classifiers in the task of acoustic-driven pornography recognition with respect to five different evaluation metrics

Method	Precision	Recall	F-score	Accuracy	AUC
1- Coarse KNN	72.10%	84.76%	77.92%	83.00%	92%
2- Cosine KNN	85.02%	83.44%	84.22%	86.70%	94%
3- Cubic KNN	83.04%	85.55%	84.27%	87.10%	94%
4- Fine KNN	83.22%	86.04%	84.61%	87.40%	87%
5- Medium KNN	84.32%	86.09%	85.20%	87.80%	95%
6- Weighted KNN	83.68%	85.79%	84.72%	88.40%	87%

As can be observed in the table, three out of the five best performances were achieved by the Medium KNN classifier in terms of recall, F-score, and AUC at 86.09%, 85.20%, and 95.00%, respectively. Meanwhile, the best precision and accuracy rates were obtained by Cosine KNN and Weighted KNN, respectively. The importance of applying a voting schema based on class weights is highlighted by the results of the Weighted KNN.

Although the best accuracy was achieved by the Weighted KNN at 88.40%, the other three classifiers, Medium, Fine, and Cubic, successfully achieved comparable performance with marginal differences at 0.6%, 1.0%, and 1.3%, respectively, placing them at the second, third, and fourth places in terms of accuracy.

Since the difference between the best recall rate and the one obtained by Fine KNN was only 0.05%, one may consider their performances to be the same. The AUC values, as more appropriate indicators of performance in the binary classification problems, highlight the fact that all three variants of the KNN classifiers, namely, Cosine, Cubic, and Medium, could result in the same acoustic pornography recognition rates with marginal differences. In other words, the AUC values indicate that, regardless of the threshold set for classification, all three variants yielded almost the same binary classification performance in the case of acoustic pornography recognition.

6. CONCLUSION

In this research, we used acoustic information extracted from video clips in order to train different supervised classification models and test the feasibility of acoustic-driven features in the task of pornography recognition. More specifically, two types of features, pitch and MFCC, were employed to construct acoustic representations of the audio tracks.

We constructed a new audio dataset of pornography soundtracks comprising two sets of training and testing partitions. After conducting multiple experiments, the best performance enhancement in terms of recall, F-score, and AUC was achieved by the Medium KNN, and the highest recognition rates for precision and accuracy were obtained by Cosine KNN and Weighted KNN, respectively.

In future works, we intend to extend our research by investigating the effects of other pitch-based feature descriptor algorithms, such as those reported in studies by Drugman and Alwan (2011), Gonzalez and Brookes (2011), Hermes (1988), and Noll (1967). We will also explore the performance of different supervised and unsupervised learning models on a larger pornography

audio dataset.

7. ACKNOWLEDGEMENT

This research was fully funded by TELEKOM Malaysia Research and Development (TM R&D).

8. REFERENCES

Atal, B.S., 1972. Automatic Speaker Recognition based on Pitch Contours. *The Journal of the Acoustical Society of America*, Volume 52(6B), pp. 1687–1697

Caetano, C., Avila, S., Schwartz, W.R., Guimarães, S.J.F., Araújo, A. de A., 2016. A Mid-level Video Representation based on Binary Descriptors: A Case Study for Pornography Detection. *Neurocomputing*, Volume 213, pp. 102–114

Dalal, N., Triggs, B., Europe, D., 2005. Histograms of Oriented Gradients for Human Detection. *In:* IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05), 20-25 June 2005

Drugman, T., Alwan, A., 2011. Joint Robust Voicing Detection and Pitch Estimation based on Residual Harmonics. *In:* Twelfth Annual Conference of the International Speech Communication Association, 27-31 August 2011

Foggia, P., Petkov, N., Saggese, A., Strisciuglio, N., Vento, M., 2015. Reliable Detection of Audio Events in Highly Noisy Environments. *Pattern Recognition Letters*, Volume 65, pp. 22–28

Foggia, P., Petkov, N., Saggese, A., Strisciuglio, N., Vento, M., 2016. Audio Surveillance of Roads: A System for Detecting Anomalous Sounds. *IEEE Transactions on Intelligent Transportation Systems*, Volume 17(1), pp. 279–288

Gemmeke, J.F., Ellis, D.P.W., Freedman, D., Jansen, A., Lawrence, W., Moore, R.C., Plakal, M., Ritter, M., 2017. Audio Set: An Ontology and Human-labeled Dataset for Audio Events. *In:* International Conference on Acoustics, Speech, and Signal Processing, pp. 776–780

Geng, Z., Zhuo, L., Zhang, J., Li, X., 2016. A Comparative Study of Local Feature Extraction Algorithms for Web Pornographic Image Recognition. *In:* Proceedings of 2015 IEEE International Conference on Progress in Informatics and Computing, PIC 2015, pp. 87–92

Gonzalez, S., Brookes, M., 2011. A Pitch Estimation Filter Robust to High Levels of Noise (PEFAC). *In:* European Signal Processing Conference, (Eusipco), 29 August - 2 September 2011

Gupta, M., Bhaskar, D., Bera, R., 2016. Automatic Target Classification in GMTI Airborne Scenario. *International Journal of Technology*, Volume 7(5), pp. 840–848

Hasan, R., Jamil, M., Rabbani, G., Rahman, S., 2004. Speaker Identification using Mel Frequency Cepstral Coefficients. *In:* Proceedings of the 3rd International Conference on Electrical & Computer Engineering (ICECE 2004), December 2004, pp. 28–30

Hermes, D.J., 1988. Measurement of Pitch by Subharmonic Summation. *The Journal of the Acoustical Society of America*, Volume 83(1), pp. 257–264

Jiang, Y.G., Bhattacharya, S., Chang, S.F., Shah, M., 2013. High-level Event Recognition in Unconstrained Videos. *International Journal of Multimedia Information Retrieval*, Volume 2(2), pp. 73–101

Jin, X., Wang, Y., Tan, X., 2018. Pornographic Image Recognition via Weighted Multiple Instance Learning. *IEEE Transactions on Cybernetics*, Volume 49(12), pp. 4412–4420

Lopes, A.P.B., De Avila, S.E.F., Peixoto, A.N.A., Oliveira, R.S., Coelho, M.D.M., Araújo, A.D.A., 2009. Nude Detection in Video using Bag-of-visual-features. *In:* Proceedings of SIBGRAPI 2009, 22nd Brazilian Symposium on Computer Graphics and Image Processing, pp. 224–231

Lowe, D.G., 1999. Object Recognition from Local Scale-invariant Features. *In:* Proceedings of

the Seventh IEEE International Conference on Computer Vision, Volume 2, pp. 1150–1157

Mesaros, A., Heittola, T., Virtanen, T., 2016. TUT Database for Acoustic Scene Classification and Sound Event Detection. *In:* European Signal Processing Conference (EUSIPCO), November 2016, pp. 1128–1132

More, M.D., Souza, D.M., Barros, R.C., 2018. Seamless Nudity Censorship: An Image-to-Image Translation Approach based on Adversarial Training. *In:* IEEE International Joint Conference on Neural Networks (IJCNN)

Moreira, D., Avila, S., Perez, M., Moraes, D., Testoni, V., Valle, E., Goldenstein, S., Rocha, A., 2016. Pornography Classification: The Hidden Clues in Video Space–time. *Forensic Science International*, Volume 268, pp. 46–61

Naik, S., Metkewar, P., 2015. Recognizing Offline Handwritten Mathematical Expressions (ME) based on a Predictive Approach of Segmentation using K-NN Classification. *International Journal of Technology*, Volume 6(3), pp. 345–354

Nian, F., Li, T., Wang, Y., Xu, M., Wu, J., 2016. Pornographic Image Detection Utilizing Deep Convolutional Neural Networks. *Neurocomputing*, Volume 210, pp. 283–293

Noll, A.M., 1967. Cepstrum Pitch Determination. *The Journal of the Acoustical Society of America*, Volume 41, pp. 293–309

Nurhadiyatna, A., Cahyadi, S., Damatraseta, F., Rianto, Y., 2018. Adult Content Classification through Deep Convolution Neural Network. *In:* Proceedings of the 2017 International Conference on Computer, Control, Informatics and Its Applications: Emerging Trends In Computational Science and Engineering, IC3INA 2017, January 2018, pp. 106–110

Piczak, K.J., 2015. ESC: Dataset for Environmental Sound Classification. *In:* Proceedings of the 23rd ACM International Conference on Multimedia, pp. 1015–1018

Pieropan, A., Salvi, G., Pauwels, K., Kjellstrom, H., 2014. Audio-visual Classification and Detection of Human Manipulation Actions. *In:* IEEE International Conference on Intelligent Robots and Systems (IROS), pp. 3045–3052

Rabiner, L.R., Schafer, R.W., 2011. *Theory and Applications of Digital Speech Processing.* Pearson, Upper Saddle River, NJ

Salamon, J., Bello, J.P., 2015. Unsupervised Feature Learning for Urban Sound Classification. *In:* Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), August 2015, pp. 171–175

Shen, R., Zou, F., Song, J., Yan, K., Zhou, K., 2018. EFUI: An Ensemble Framework using Uncertain Inference for Pornographic Image Recognition. *Neurocomputing*, Volume 322, pp. 166–176

Shi, Z., Han, J., Zheng, T., Li, J., 2013. Identification of Objectionable Audio Segments based on Pseudo and Heterogeneous Mixture Models. *IEEE Transactions on Audio, Speech and Language Processing*, Volume 21(3), pp. 611–623

Snoek, C.G.M., Worring, M., 2007. Concept-based Video Retrieval. *Foundations and Trends® in Information Retrieval*, Volume 2(4), pp. 215–322

Snoek, C.G.M., Worring, M., Smeulders, A.W.M., 2005. Early Versus Late Fusion in Semantic Video Analysis. *In:* Proceedings of the 13th Annual ACM International Conference on Multimedia, MULTIMEDIA '05, pp. 399–402

Stowell, D., Giannoulis, D., Benetos, E., Lagrange, M., Plumbley, M.D., 2015. Detection and Classification of Acoustic Scenes and Events. *IEEE Transactions on Multimedia*, Volume 17(10), pp. 1733–1746

Zhou, K., Zhuo, L., Geng, Z., Zhang, J., Li, X.G., 2016. Convolutional Neural Networks Based Pornographic Image Classification. *In:* Proceedings of the 2016 IEEE 2nd International Conference on Multimedia Big Data, BigMM 2016, pp. 206–209

Zhou, W., Ahrary, A., Kamata, S.I., 2012. Image Description with Local Patterns: An Application to Face Recognition. *IEICE Transactions on Information and Systems*, Volume E95-D(5),

pp. 1494–1505
Zuo, H., Wu, O., Hu, W., Xu, B., 2008. Recognition of Blue Movies by Fusion of Audio and Video. *In:* Proceedings of the 2008 IEEE International Conference on Multimedia and Expo, ICME 2008, pp. 37–40