OUT-OF-SCENE AV DATA DETECTION

Danil Korchagin
Idiap Research Institute
P.O. Box 592, CH-1920 Martigny, Switzerland

ABSTRACT

In this paper, we propose a new approach for the automatic audio-based out-of-scene detection of audio-visual data, recorded by different cameras, camcorders or mobile phones during social events. All recorded data is clustered to out-of-scene and in-scene datasets based on confidence estimation of cepstral pattern matching with a common master track of the event, recorded by a reference camera. The core of the algorithm is based on perceptual time-frequency analysis and confidence measure based on distance distribution variance. The results show correct clustering in 100% of cases for a real life dataset and surpass the performance of cross correlation while keeping lower system requirements.

KEYWORDS

Time-frequency analysis, pattern matching, confidence estimation.

1. INTRODUCTION

The TA2 project (Together Anywhere, Together Anytime) is concerned with investigation of how multimedia devices can be introduced into a family scenario to break down technology and distance barriers. Many of our enduring experiences, holidays, celebrations, moments of fun and laughter are framed as family events. Modern media and communications serve individuals best, with phones, computers and electronic devices tending to be individually owned and providing individual experiences. In this sense, we are interested in the use of consumer level multimedia devices in novel application scenarios which can help to nurture family-to-family relationships.

One generic scenario is the use of multiple capture devices in a single room by a group of socially together families. The purpose of each of the video collections is primarily personal: each family is interested in capturing their own favorite, but also enough context information from the event to provide some background. Not everybody films all the time. Each family is interested in creating a video fragment for the personal family archives, plus short clips that can be sent to family members at home and abroad who were not present. All of the people are used to multi-camera concert videos: they would like to make use of the material of others when it is appropriate – either to show the shot they missed or to provide a more compelling video by having multiple views.

The present investigation concerns the possibility of using multiple consumer level video cameras with automatic separation of different events. Consumer level devices, however, do not normally provide such capabilities. Many video clustering techniques are based on the video signal analysis (Agnihotri 2000, Xie 2008). Nevertheless, if the devices are hand-held, we cannot rely in any predictable sense on the video signal. This leaves us with the audio signal (Dannenberg 2003) from which to infer clustering information.

In this study, we were provided with a few reference signals from fixed cameras that recorded the whole scene. We were also provided with several auxiliary signals from hand-held cameras that recorded parts of different scenes. If we could show that the auxiliary signals could be reliably clustered to in-scene and out-of-scene datasets with respect to the reference signal, then the project could use open datasets and let the system automatically separate different events. If it were too error-prone or computationally onerous, then only closed dataset would have to be considered.

2. OUT-OF-SCENE DETECTION

Consider a simple high school concert event. The duration of the corresponding master track can easily be of the order of a small number of hours. This in turn corresponds to a large quantity of raw audio data (stereo at 48 kHz). It is normal in such situations to decrease the search space, retaining only useful information for in-scene and out-of-scene clustering. Accordingly, we assume from the outset that raw Pulse Code Modulation (PCM) audio data is both too voluminous and too noisy to produce reliable decision. We suppose that good results might be obtained by lower resolution features such as frame energy and cepstra.

Given that our broader application is expected to include Automatic Speech Recognition (ASR), the pre-processing takes the form of a standard feature extraction chain used in ASR. In our work we use Mel Frequency Cepstral Coefficients (MFCC) (Mermelstein 1976) with a 10 ms frame rate. MFCC is a perceptually motivated spectrum representation that is widely used not only in speech recognition but also for music modelling (Logan 2000). Such pre-processing includes energy-like features (actually the zero'th cepstral coefficient) along with cepstra representing the general spectral shape.

We assume that test samples are relatively short, thus we can ignore the clock skew problem between test and reference (i.e., there is almost zero skew due to unsynchronised clocking of different devices). Presumably in some cases for long recordings the two could become misaligned, in which case additional techniques such as dynamic time warping (Hu 2003) should be taken into account during the matching process. We consider two operating modes, one the well-known cross correlation and the other pattern matching based on ASR-related features.

2.1 Cross correlation

Cross correlation is a measure of similarity of two waveforms as a function of a time-lag applied to one of them. It can be used to search a long duration signal for a shorter. If h_i and g_j are the raw test and reference signals respectively, and h^* is the complex conjugate of h, then C_i^j, the confidence of the i'th test clip belongs to j'th reference signal, is given by:

$$C_i^j = \frac{1}{40\sigma} \max_n \left(\sum_m h_i^*(m) g_j(n+m) \right),$$

where σ is the standard deviation of the cross correlation.

Regardless of the simplicity of implementation, standard cross correlation cannot be implied by our scenario as it is computationally onerous (several days per clip on an Intel Core 2 CPU 6700 2.66GHz), nevertheless this can be resolved by the convolution theorem and the fast Fourier transform, also known as fast cross correlation:

$$C_i^j = \frac{1}{40\sigma} \max \left(F^{-1} \left(\left(F\{h_i\} \right)^* \cdot F\{g_j\} \right) \right)$$

In the above formulation, the parameters are as before, except F denotes the fast Fourier transform. An asterisk again indicates the complex conjugate. The processing time for fast cross correlation takes only 70 seconds per clip, though it requires much more RAM (3 GB versus 100 MB).

2.2 Pattern matching

Audio is resampled to mono 16 kHz and pre-emphasised to flatten the spectral shape. A 256 point Discrete Fourier Transform (DFT) is performed in steps of 10 ms and squared to give the power spectrum. The resulting 129 unique bins are then decimated using a filter-bank of 23 overlapping triangular filters equally spaced on the mel-scale. The mel-scale corresponds roughly to the response of the human ear. A logarithm and DFT then yield the mel-cepstrum, which is truncated, retaining the lower 13 dimensions. This truncation retains spectral shape and discards excitation frequency. Next, Cepstral Mean Normalisation (CMN) is performed by subtracting from each cepstral vector the mean of the vectors of the preceding (approximately) half second. This has the effect of removing convolutional channel effects. Finally, if the norm of a vector of the 13 mean normalised cepstral coefficients is higher than 1, then the vector is normalised in Euclidean space. This gives us the reduced variance of the search distance space.

Pattern matching based on the above features is performed by searching for a best distance in Euclidean space between the test time-quefrency matrix (corresponding to a test clip) and the master time-quefrency matrix in steps of 10 ms. If V_i is the i'th test matrix, M^j is the j'th master matrix and M_p^j is the sub-matrix of the master matrix, shifted from the beginning by $10p$ ms, then C_i^j, the confidence of the i'th test clip belongs to j'th reference signal, is given by:

$$C_i^j = \frac{\left|\mathrm{E}(d(M_p^j, V_i)) - \min_p(d(M_p^j, V_i))\right|}{4 \cdot \left|\mathrm{E}(d(M_p^j, V_i)) - \max_p(d(M_p^j, V_i))\right|} - \frac{20}{N_i},$$

where d is Euclidean metric, E is expectation and N_i is the number of frames inside test matrix (used to decrease false in-scene detections for short segments).

2.3 Experimental results

All results presented in this paper were achieved on a real life dataset of 125 recordings (2 master tracks + 123 test recordings, table 1).

Table 1. Experimental dataset

Source	Length range	Audio specification
Canon XL-G1 SD/HD	**51 min (1st master track)**	PCM, 32000Hz
Nokia N95	**21-130 s (15 recordings)**	AAC, 48000Hz
Canon FS100 mini	**15-133 s (26 recordings)**	AC3, 48000Hz
Sony DCR-PC3e	**16-695 s (15 recordings)**	PCM, 48000Hz
Sanyo Xacti HD mini	**15-250 s (25 recordings)**	AAC, 48000Hz
Siemens E71	**43 min (2nd master track)**	ARM, 8000Hz
Sony Ericsson G502	**22-39 s (7 recordings)**	AAC, 16000Hz
Sony DSC-V1	**18-90 s (35 recordings)**	MPEG, 32000Hz

The master track contents consist of a rehearsal and a concert. Experiments were conducted on an open set (123 in-scene recordings and 123 out-of-scene recordings), where only first few seconds (defined by the experiments) of each recording were used for the decision.

In figure 1 we illustrate how the length of the test segments influences in-scene confidence estimation (the means of all experiments are shown, 0 corresponds to 0% of subjective confidence, 1 corresponds to 100% of subjective confidence). The corresponding standard deviation on average is 0.15 for in-scene cepstral measures, 0.20 for in-scene cross correlation, 0.04 for out-of-scene measures.

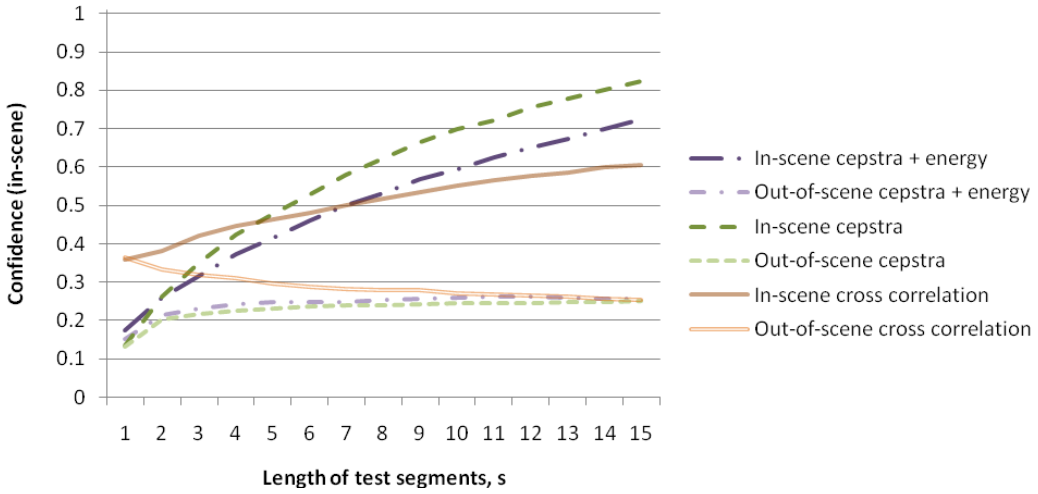

Figure 1. In-scene confidence versus test segment length

It is clearly visible that the confidence for in-scene segments increases with increasing the length of test segments. The in-scene confidence for out-of-scene segments slightly increases for the cepstral measures and decreases for the cross correlation with increasing the length of test segments. This is due to adaptive minimization of false in-scene detections used within cepstral measures when there is not enough information for confident decision (e.g., for short segments). The corresponding performance (the number of correctly clustered clips divided by the total number of test clips) versus confidence threshold is shown in figure 2 (for maximum length thresholded at 60 seconds).

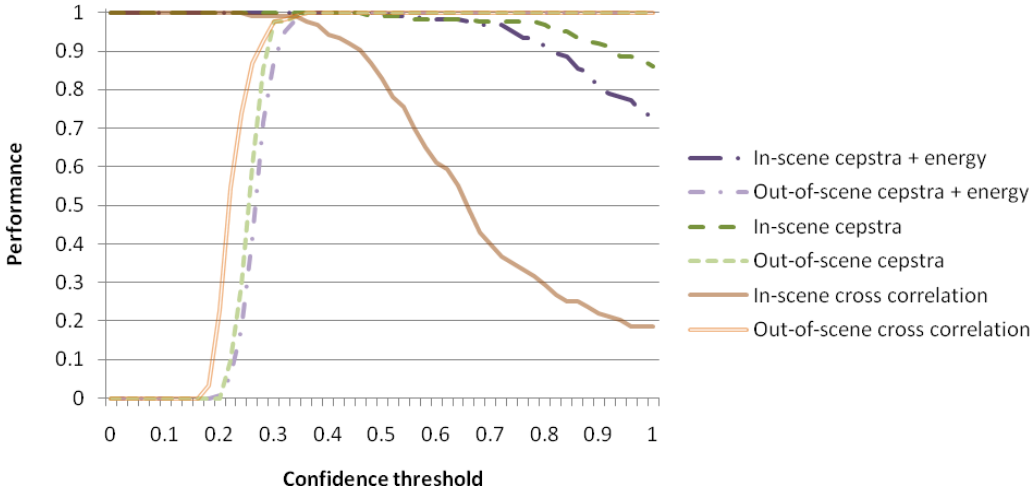

Figure 2. Performance versus confidence threshold

The corresponding performance is lower when the energy is considered (dash dot and long dash dot lines) due to the increased variance of the distance space. From the figure 2 we can estimate the optimal fixed thresholds based on intersections of the in-scene and out-of-scene performance for cepstral (optimal fixed confidence threshold is 0.36) and cross correlation (optimal fixed confidence threshold is 0.3) modes. The use of lower thresholds increases the probability of false in-scene detection. There is also a strong dependency on the length of test recordings, which is shown in figure 3. The performance increases and, for recordings longer than 20 s, 100% performance is achievable on the described dataset for the proposed approach versus 99% for cross correlation. The performance for shorter segments is lower due the real world variability of the data (noise, reverberation, non-stationarity of cameras, inter-microphone variability, inter-codec variability, etc).

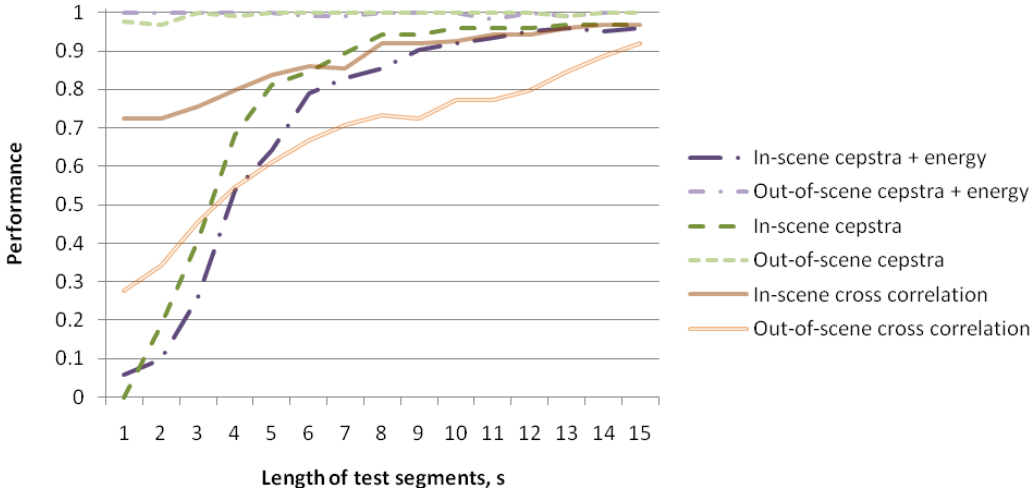

Figure 3. Performance versus test segment length

The processing time (on Intel Core 2 CPU 6700 2.66GHz) for the proposed algorithm without multi-core optimisation is on average 8 seconds per 15 seconds test file over each 10 minutes of master track. The computational efficiency of proposed approach is better than fast cross correlation (41 seconds versus 70 seconds for fast cross correlation per clip) and memory requirement is about 15% of the size of reference signal (15 MB versus 3 GB for fast cross correlation).

3. CONCLUSION

We have shown that multiple AV signals can be clustered to an acceptable accuracy using audio features typical of ASR applications. We found that the energy of the signal is not good for clustering, but that reliable clustering can be inferred from a small number of normalised cepstra. We have estimated that results surpass the performance of fast cross correlation, while requiring less resources. The achieved results allow us in the future to enlarge an application domain by enabling automatic unsupervised multimedia structuring and new representation of the media for graphical user interfaces.

ACKNOWLEDGEMENT

The research leading to these results has received funding from the European Community's Seventh Framework Programme ICT Integrating Project “Together Anywhere, Together Anytime” (TA2, FP7/2007-2013) under grant agreement no. ICT-2007-214793. I should like to extend my gratitude to Philip N. Garner and John Dines for their valuable help at various stages of this work and provision of the feature extraction software. I am grateful to British Telecom for provision of the real life dataset.

REFERENCES

Agnihotri, L. and Dimitr, N., 2000. Video Clustering Using Super Histograms in Large Archives. *LNCS, Advances in Visual Information Systems.* Springer Berlin / Heidelberg, Germany, Vol. 1929/2000, pp. 255-268.

Dannenberg, R.B. and Hu, N., 2003. Polyphonic Audio Matching for Score Following and Intelligent Audio Editors. *Proceedings of the 2003 International Computer Music Conference.* San Francisco, USA, pp. 27-34.

Hu, N. et al., 2003. Polyphonic Audio Matching and Alignment for Music Retrieval. *In 2003 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics.* New York, USA, pp. 185-188.

Integrating Project within the European Research Programme 7, 2008. *Together Anywhere, Together Anytime.* http://www.ta2-project.eu/ .

Logan, B., 2000. Mel Frequency Cepstral Coefficients for Music Modeling. *Proceedings of the 1st International Symposium on Music Information Retrieval.* Plymouth, USA.

Mermelstein, P., 1976. Distance Measures for Speech Recognition, Psychological and Instrumental. *In Pattern Recognition and Artificial Intelligence.* C. H. Chen, Ed., Academic, New York, pp. 374-388.

Xie, X.-N. and Wu, F., 2008. Automatic Video Summarization by Affinity Propagation Clustering and Semantic Content Mining. *Proceedings of the 2008 International Symposium on Electronic Commerce and Security*. Guangzhou City, China, Vol. 00, pp. 203-208.