CAST a database: Rapid targeted large-scale big data acquisition via small-world modelling of social media platforms

Shahin Amiriparian*†, Sergey Pugachevskiy*, Nicholas Cummins*, Simone Hantke*†, Jouni Pohjalainen*,
Gil Keren* and Björn Schuller*‡
**Chair of Complex & Intelligent Systems, University of Passau, Germany*
†*Machine Intelligence & Signal Processing Group, Technische Universität München, Germany*
‡*Machine Learning Group, Imperial College London, U.K.*
Email: shahin.amiriparian@tum.de

Abstract—**The adage that there is no data like more data is not new in affective computing; however, with recent advances in deep learning technologies, such as end-to-end learning, the need for extracting big data is greater than ever. Multimedia resources available on social media represent a wealth of data more than large enough to satisfy this need. However, an often prohibitive amount of effort has been required to source and label such instances. As a solution, we introduce Cost-efficient Audio-visual Acquisition via Social-media Small-world Targeting (CAS^{2}T) for efficient large-scale big data collection from online social media platforms. Our system is based on a unique combination of small-world modelling, unsupervised audio analysis, and semi-supervised active learning. Such an approach facilitates rapid training on entirely new tasks sourced in their entirety from social multimedia. We demonstrate the high capability of our methodology via collection of original datasets containing a range of naturalistic, in-the-wild examples of human behaviours.**

1. Introduction

Contemporary deep topologies such as end-to-end learning, reinforcement learning, and representation learning require significantly more data than conventional machine learning approaches [1]. Therefore, research fields that relies heavily on data that is laborious and costly to collect and annotate, such as affective and behavioural computing, struggle to procure enough reliably labelled instances to adequately train such systems [2].

However, the good news for researchers and analysts who would like to exploit these new technologies is that the required data is out there. Information technology companies such as Google, Apple, Microsoft, and Facebook collect and maintain data in exabyte proportions [3]. Importantly, a large proportion of data has been made publicly available with very few restrictions limiting it's collection.

One such resource is social media and video sharing websites. For example, it was estimated that in July 2015 more than 400 hours of video data were uploaded to the popular video sharing website YouTube every minute [4]. This vast and growing amount of in-the-wild multimedia material publicly available online, including produced content and non-acted personal home videos, represents an untapped wealth of data for research purposes.

Equally important as large and available data resources, is the advent of machine learning paradigms which enable the labelling of large datasets with minimal human assistance [5], [6]. Strategies like semi-supervised active learning use confidence values from a machine learning algorithm to determine whether to keep the automatically assigned label or to request a human to label the instance in question. Such paradigms have been used to aid multimodal emotion recognitionand sound classification tasks [6], [7].

Moreover, thanks to recent advances in crowdsourcing platforms, when human annotation is required, the labels can be obtained efficiently with low financial overheads. Crowdsourcing is the utilisation of a large group of non-experts to perform a common task; the underlying assumption being the collective opinion of this large group is quicker and less demanding to obtain than that of a smaller group of trained experts. Further, this collective opinion has been shown to be as high a quality as ones determined by small groups of experts, at a fraction of the cost [8], [9]. Crowdsourcing has been successfully used in a range of affective computing applications [10], [11], [12].

While the data and the basic tools needed to generate big datasets have been available, to the best of the authors knowledge, no system exists which combines them to efficiently source multimedia data from internet resources. In this regard, we herein introduce our COST-EFFICIENT AUDIO-VISUAL ACQUISITION VIA SOCIAL-MEDIA SMALL-WORLD TARGETING (CAS^{2}T) system that is purpose-built for enabling fast, efficient and reliable audio-visual data collection directly from social media platforms in a cost-efficient manner.

Our approach combines complex systems theory (Section 2.1), unsupervised audio analysis (Section 2.2), semi-supervised active learning (Section 2.3) and crowdsourcing (Section 2.4). In heralding this solution, we collect six new audio databases (Section 3) and demonstrate, via a set of classification experiments, the effectiveness of our system (Section 4). Finally, our conclusions and future work plans are given in (Section 5).

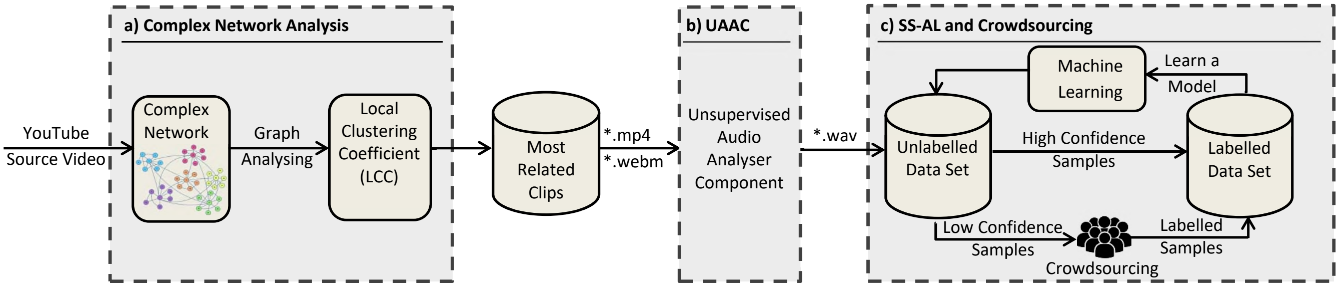

Figure 1: Block diagram of CAS^{2}T. A network of interwoven YouTube videos is analysed by complex systems components (a) in order to identify an initial set of 'related' multimedia clips available on social media, specified by the source video. These clips are then passed to an *Unsupervised Audio Analyser Component* (UAAC) which performs unsupervised event detection to isolate audio instances of potential interest (b). We then use a mix of *semi-supervised active learning* (SS-AL) and *crowdsourcing* (c) to label the data with minimal human intervention. Samples with low classification certainty (low confidence samples) are selected to be sent for human annotation, and the instances with high classification certainty (high confidence samples) are directly added to training data set with labels automatically determined by the machine annotator. $.mp4$ and $.webm$ are extensions for video files extracted from YouTube; $.wav$ is an extension of audio files.

2. Multimedia database generation

While online multimedia archives contain a wealth of data, its practical application in training machine learning systems is hindered by three obstacles: 1) finding the relevant recordings; 2) segmenting these into meaningful and coherent segments; and 3) reliably labelling the segments so that they can be useful in machine learning. In the following subsections, we describe our solutions to these problems.

2.1. Complex Network Analyser Component

The role of the *Complex Network Analyser Component* (CNAC) is to enable fast identification of 'related' multimedia data from online resources. It is currently implemented to work with YouTube – the world's largest video sharing website. YouTube has one of the world's largest recommendation systems [13], which offers a viewer suggestions as to the next video they should watch. Whilst the exact workings of this system are not publicly available, it is known to be based on features including: the number of video views; video title, description, and associated metadata; search query tokens; viewer demographics; and the video rankings (i. e. number of likes and dislikes) [13].

The CNAC operates under the assumption that the content of recommended videos is, in general, highly similar (i. e. related) to the original video and exploits the complex networks of interconnections generated by YouTube's recommendation system by modelling them as graphs with small-world properties [14]. Given an initial source video link (cf. 'YouTube Source Video' in Figure 1), the CNAC uses the YouTube Data API to build an undirected graph G of the videos most highly recommended to the viewer. The *vertices* of G represent videos that are considered to be potentially related to the topic of interest (source video), and the *edges* correspond to the recommendations between videos.

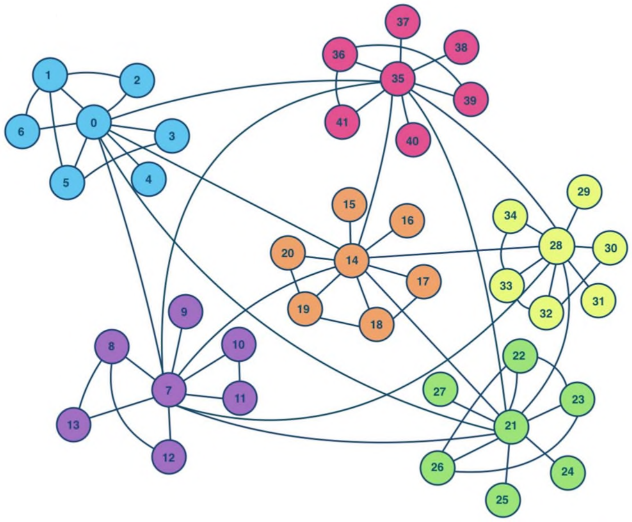

Figure 2: An illustration of the graph G for six searched subjects. YouTube clips with high Local Clustering Coefficient (LCC) are more likely to build a clique, as shown by the different coloured nodes in the graph, and are related to a specific topic.

The graph G can therefore be thought of as a mapping of YouTube's recommendation space in relation to the source video. To the best of the authors' knowledge, such a mapping has not been previously realised. This mapping is required to reveal the extent of mutual relationships between related videos of interest. Within G we assume the videos (vertices) with a greater relation to the topic of interest to have a higher number of connections (edges) with other vertices. Accordingly, we use *Local Clustering Coefficient* (LCC) algorithms to identify highly related videos [14], [15], [16].

The LCC of a video v_i in G quantifies how close its neighbours are to being a clique (complete graph) and how likely they are part of larger highly connected videos groups. This can be calculated via

$$C_{v_i} = \frac{2n}{k_{v_i}(k_{v_i} - 1)}$$

where C_{v_i} is the LCC for v_i, n is the number of edges that actually pass between the neighbours of v_i, and k_{v_i} is the number of neighbours of v_i [14]. Highly related videos in G, videos with a high number of edges (i. e. recommendations), will have a high LCC value and conversely, unrelated videos will have a low LCC. As illustrated in Figure 2, the CNAC uses the LCC to locate highly mutually related content groups in G, which can be downloaded and then sent to the UAAC for further processing.

2.2. Unsupervised Audio Analyser Component

The goal of the *Unsupervised Audio Analyser Component* (UAAC) is to extract coherent, meaningful segments from the collected videos. In achieving this, instances can be processed by one of two different approaches: *energy-based* and *spectrum-based*. The former is used for videos in which the target class presents itself as isolated audio events of approximately known duration against a relatively quiet background (i.e. the signal-to-noise ratio is high during the events) and not many other prominent audio sources are present.

The spectrum-based approach is used when the target audio events are not necessarily prominent in terms of energy but are still expected to be distinguishable based on their spectra, i.e. in order to detect events based on their spectral rather than temporal coherence. Both approaches share a similar post-processing stage to produce segments of desired length.

Denoting the logarithmic energy of the nth short-time audio frame as E_n, the energy-based detector computes for each frame the statistic $E_{n,\beta} - E_{n,\alpha}$, where the two terms are lowpass-filtered versions of E_n produced using different memory coefficients so that $\alpha > \beta$, and subtracts the longer-term average (assumed to be the slowly changing noise floor level) from the shorter-term average (which may be higher due to energetic events). Unsupervised thresholding is used to extract frames with high values for post-processing.

The spectrum-based detector generates an unsupervised classification of spectral representations of each frame – here 50-dimensional linear-frequency cepstral vectors – and selects the frames belonging to the J out of K clusters with the highest frame energy for post-processing. It thus makes use of both mutual spectral similarity and signal energy.

The post-processing connects all segments that are separated by less than t_{skip} seconds and subsequently discards all segments whose length is smaller than $t_{\min}$ seconds or larger than $t_{\max}$ seconds.

When gathering our dataset (cf. Section 3), we used the following UAAC parameter values (tuned empirically): the audio frame length was 30 ms spaced at 10 ms frame shift intervals; $\alpha = 0.999$, β =0.95, $t_{\text{skip}} = 0.2$ s, $t_{\min} = 0.5$ s, and $t_{\max} = 10$ s for the energy-based detector, which is used to detect short audio events, and $K = 128$, $J = 116$, $t_{\text{skip}} = 0.25$ s, $t_{\min} = 2.0$ s, and $t_{\max} = 30$ s for the spectrum-based detector, which is used to detect speaking styles.

2.3. Semi-Supervised Active Learning

Semi-Supervised Active Learning (SS-AL) has consistently been shown to be a reliable method to reduce human efforts when annotating data [7]. This approach shares labelling work between humans and a (pre-trained) machine learning system in an iterative manner. The SS-AL approach used in our solution is based on the *Semi-Supervised Active Learning in a pool-based scenario* paradigm presented in [7]. The confidence thresholds $C_{low} = 0.15$ and $C_{high} = 0.85$ are determined by the performance of the SVM classifier. All instances with the confidence values Cs $> C_{low}(0.15)$ and Cs $\leq C_{high}(0.85)$ will remain in the pool of unlabelled data $\mathcal{U}$. An overview of this algorithm is given in Algorithm 1.

Algorithm 1 Semi-supervised Active Learning

```
Given:
  • $\mathcal{L}$: a small set of labelled data
  • $\mathcal{U}$: a large pool of unlabelled data
  • $\mathcal{M}$: an initial model trained by $\mathcal{L}$
  • $C_{low} = 0.15$, $C_{high} = 0.85$: classifier's confidence thresholds
repeat
  • Using $\mathcal{M}$ to classify every instance in $\mathcal{U}$, and calculate the corresponding classifier's confidence value $C$
  • Choose instances with $C$s $< C_{low}$ from $\mathcal{U}$, and send them for annotation to manual annotation
  • Refer to the new labelled set as $\mathcal{U}^o_{new}$
  • $\mathcal{L} = \mathcal{L} \cup \mathcal{U}^o_{new}$ and $\mathcal{U} = \mathcal{U} - \mathcal{U}^o_{new}$
  • Re-train $\mathcal{M}$ using the new $\mathcal{L}$
  • Choose instances with $C$s $\geq C_{high}$ from $\mathcal{U}$, and add the corresponding predicted labels
  • Refer to the new machine-labelled set as $\mathcal{U}^m_{new}$
  • $\mathcal{L} = \mathcal{L} \cup \mathcal{U}^m_{new}$ and $\mathcal{U} = \mathcal{U} - \mathcal{U}^m_{new}$
  • Re-train $\mathcal{M}$ using the new $\mathcal{L}$
until there is no data in the pool predicted as belonging to the target class OR model training converges OR manual annotation is not possible
```

2.4. Harnessing the Crowd

All manual labelling is performed using the crowdsourcing-based data recording and annotation

TABLE 1: *Specifications of the samples in each data set. l_{total}: total length of the data set; l_{min} and l_{max}: the minimum and maximum sample lengths; σ: standard deviation; n_{tg}/n_{sp}: number of all target and speech samples in each set*

	Train				***Evaluation***			
Tasks	l_{total}	l_{min}/l_{max}	σ	n_{tgt}/n_{sp}	l_{total}	l_{min}/l_{max}	σ	n_{tg}/n_{sp}
Freezing	$75.9\,m$	$2.0\,s/29.4\,s$	$5.8\,s$	409/205	$22.4\,m$	$2.0\,s/28.6\,s$	$5.9\,s$	91/80
Intoxication	$139.7\,m$	$2.0\,s/29.9\,s$	$6.5\,s$	514/555	$16.7\,m$	$2.0\,s/24.8\,s$	$5.3\,s$	97/55
Screaming	$53.6\,m$	$2.0\,s/29.9\,s$	$7.6\,s$	203/172	$22.0\,m$	$2.1\,s/29.9\,s$	$5.5\,s$	111/78
Threatening	$106.6\,m$	$2.0\,s/29.8\,s$	$7.4\,s$	559/93	$45.8\,m$	$2.0\,s/29.2\,s$	$5.2\,s$	163/278
Coughing	$94.3\,m$	$0.5\,s/28.8\,s$	$3.5\,s$	1 553/535	$63.9\,m$	$0.5\,s/23.2\,s$	$2.7\,s$	1 080/491
Sneezing	$6.7\,m$	$0.5\,s/8.0\,s$	$1.3\,s$	114/124	$9.2\,m$	$0.5\,s/9.3\,s$	$1.4\,s$	150/141

platform *iHEARu-PLAY* [8]. This platform provides volunteers a game-like environment for recording and annotating speech, features a scoring system, player ranks, multiple leaderboards, and unlockable badges. Besides this gamification concept, the platform also has different mechanisms to ensure the reliability of the labelled data by the players. A quality control system and methods of identifying the within-user-agreement are integrated into iHEARu-PLAY. Each player starts with the level of trustability of 100%. In addition to the normal questions, some *control questions* and *consistency questions* can be interspersed in order to validate the annotators' concentration during the game.

Consistency questions are repeated questions on the same file within the given task and each given answer is internally compared with the previous answer(s). Control questions are automatically mixed into the stream of regular tasks and include definitely wrong answer options that can be chosen by the user. Users' answers to these questions influence their *Trustability Score*. Giving 'inconsistent' or 'wrong' answers decreases their Trustability whilst a 'correct' answer increases or maintains it. These scores can be used in a variety of methods in order to improve overall annotation quality.

When gathering our datasets, 14 player produced annotations. Since metadata is given voluntarily, we received the data from 11 players (5 female, 6 male) aged 24 to 35 (mean: 28, standard deviation: 3.1 years) for our six audio tasks.

3. Creating the Databases

For our classification experiments, we have collected six new audio databases containing different human vocalisation and speech types; namely *coughing*, *sneezing*, *freezing* – speech produced by an individual shivering with cold, *intoxication* – speech produced under the influence of drugs or alcohol, *screaming*, and *threatening* – speech perceived by our annotators to be of a threatening manner.

These datasets are based on the concept of acoustic surveillance [17]. The first two topics are related to the monitoring of everyday activity – in terms of, e.g. personal health – in common, relatively quiet environments such as home or office [17]. For these kinds of events, energy-based audio segmentation (Section 2.2) was used. The latter four topics, related to audio-based surveillance for security purposes in noisy public places, were handled using spectrum-based segmentation.

Table 1 shows an overview of the databases created. Active learning played a major role in reducing human labelling efforts when collecting this data. We used crowdsourcing to label an initial set of 3030 audio instances for use in the set of labelled data ($\mathcal{L}$) and in creating our evaluation sets (cf. Algorithm 1; Table 1). Ten rounds of active learning were applied and approximately 80% of the remaining data instances sourced were machine labelled. This represents a substantial reduction in human labelling efforts. We also used the Evaluation set to test the efficacy of our SS-AL approach. As shown in Figure 3, plotted for threatening speech, SS-AL greatly improves our system's learning results in comparison to random selection of manually labelled instances. Similar results were observed for all datasets.

4. Experiments

This section outlines the classification approaches used to generate the presented results.

4.1. Evaluation Metric

The evaluation measure chosen for the tasks is the *Unweighted Average Recall* (UAR), i. e. the mean value of recognition accuracy for each class. We use UAR as our corpora has an unbalanced class distributions (cf. Table 1).

4.2. Classification Approaches

We generated baseline results using a linear-kernel support vector machine (SVM) system trained on the *Interspeech 2009 Emotion Challenge* (IS09-emotion) [18] dataset, with the SVM cost parameter *C* optimised for each dataset.

We also use a *Bag-of-Audio-Words* (BoAW) system proposed in [19]. This system is trained using either the IS09-emotion dataset or a 39-dimensional mel-frequency cepstral coefficient (MFCC) feature representation (12-MFCC, 12-ΔMFCC, 12-ΔΔMFCC, E, ΔE, and ΔΔE, where E stands for logarithmic energy of the input speech signal). In order to learn a codebook and generate a BoAW representation, the size of the codebook S is optimised and each low-level descriptor (LLD) will be assigned to a number of its closest

TABLE 2: *Classification results of each paralinguistic task by support vector machine (SVM; linear kernel), Bag-of-Audio-Words (BoAW)+SVM, and convolutional neural network (CNN) by unweighted and weighted average (UA/WA) recall in percent. C: complexity parameter of the SVM; cw: number of closest words in the codebook to which each low-level descriptor (LLD) is then assigned; S: Size of the codebook; The chance level for each task is 50.0 % UAR.*

		SVM		*BoAW+SVM*				*CNN*
			Evaluation				**Evaluation**	**Evaluation**
Tasks	***Feature Set***	*C*	*UA (WA)*	*C*	*cw*	*S*	*UA (WA)*	*UA (WA)*
Freezing	*IS09-emotion*	10^{-6}	**70.19** (70.76)	10^{-6}	30	4 200	67.48 (69.01)	56.91 (58.48)
	MFCCs	–	–	10^{-6}	20	3 600	65.56 (66.08)	51.06 (53.22)
Intoxication	*IS09-emotion*	10^{0}	64.71 (62.50)	10^{-2}	20	3 600	**72.57** (73.03)	66.84 (69.74)
	MFCCs	–	–	10^{0}	20	3 600	66.72 (69.08)	67.54 (67.11)
Screaming	*IS09-emotion*	10^{-3}	89.21 (88.88)	10^{-5}	30	4 200	**96.98** (97.35)	89.22 (90.45)
	MFCCs	–	–	10^{0}	20	3 600	93.97 (94.71)	87.30 (88.89)
Threatening	*IS09-emotion*	10^{-5}	**73.82** (72.10)	10^{-2}	30	4 200	66.26 (72.33)	71.85 (70.75)
	MFCCs	–	–	10^{-2}	30	4 200	66.96 (72.11)	70.30 (68.41)
Coughing	*IS09-emotion*	10^{-3}	95.36 (96.37)	10^{-4}	30	4 200	96.69 (96.05)	95.44 (94.72)
	MFCCs	–	–	10^{-4}	30	4 200	**97.58** (97.52)	93.60 (92.04)
Sneezing	*IS09-emotion*	10^{-3}	79.26 (79.38)	10^{-4}	20	3 600	76.44 (76.63)	**85.16 (85.22)**
	MFCCs	–	–	10^{-3}	20	3 600	79.83 (78.46)	80.17 (80.41)

words cw. The BoAW results given were generated after finding the optimum configuration.

Moreover, we trained a convolutional neural network (CNN) architecture for all classification tasks. The network is comprised of three convolutional blocks, followed by two fully-connected layers with 500 units and a softmax layer. Each convolutional block contains two convolutional layers [20], each with 256 feature maps, the first layer having a kernel size of five time steps, and the second having a kernel size of one time step. The convolutional layers in each block were followed by max-pooling applied on groups of two time steps with a stride of two time steps. Each convolutional or fully-connected layer was followed by a rectified linear unit (ReLU) non-linearity, and Batch Normalisation [21] was applied before each non-linearity. We applied mean and variance normalisation to each MFCC feature, across each audio clip. The networks were trained on random one-second patches from the audio clips, thus augmenting the training data in a similar way to [22]. Learning was done by minimising a cross-entropy loss using stochastic gradient descent with learning rate of 0.1, using minibatches of 64 examples. At evaluation time, the predictions of each network were averaged across five one-second samples from each audio clip.

As seen in Table 2, all systems performed at well above chance level. The SVM approach achieved the best performance for the Freezing and Threatening tasks; while BoAW achieved the highest UAR for the Intoxication, Screaming, and Coughing tasks. Our neural network, with the exception of the Sneezing task, did not perform as strongly as expected; however, its performance matches the other classifiers across the tasks.

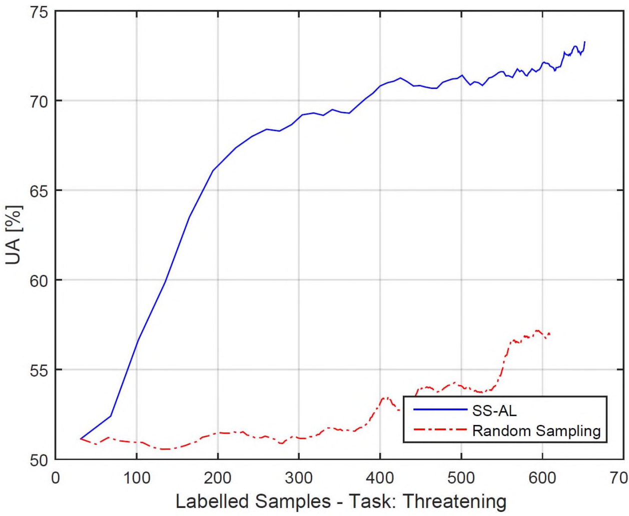

Figure 3: Learning curves of the threatening task for semi-supervised active learning (SS-AL) and random sampling. The initial training set consists of 31 samples with unweighted accuracy (UA) 51.14 %. After SS-AL, the size of the training set has increased to 652 samples with UA 73.26 %; of those added samples, 527 and 94 samples have been labelled by the machine learning system and human annotators, respectively.

5. Conclusion and Outlook

Our solution is highly effective at rapidly constructing new audio(-visual) databases or enhancing existing ones. As a consequence, data-driven approaches can benefit from the availability of additional data and achieve better performances. CAS^{2}T is a combination of complex analysis based on the small-world properties of graphs, unsupervised audio segmentation, and semi-supervised active learning. The results demonstrate that the process can be used to build datasets for use with both conventional and contemporary machine learning approaches. We have demonstrated how the vast archives of data available to us on the Internet can be harnessed for building and validating real-world acoustic surveillance systems to improve the quality of life. The inherent large *variability* as well as the sheer *volume* of online multimedia data will further enable us to develop robust systems for real-life environments by ensuring that the system evaluations are not too optimistic and that the systems will be capable of working under realistic, noisy, and unpredictable conditions.

Potential future work includes extending the system to operate on a range of social media platforms. The efficacy of our event detection may be increased by combining audio and visual information. We also plan to utilise collected data in our cross-modal representation learning research [23], [24]. Finally, the autonomy of the system can be further increased by exploring natural language processing and deep zero-resource processing techniques to enable self-gathering and self-labelling of truly large, original and in-the-wild datasets, setting the stage for the next generation of intelligent big data analytics systems.

6. Acknowledgements

This work was supported by the European Unions's Seventh Framework and Horizon 2020 Programmes under grant agreements No. 338164 (ERC StG iHEARu) and No. 688835 (RIA DE-ENIGMA).

References

[1] J. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, pp. 436–444, May 2015.

[2] B. Zhang, E. M. Provost, and G. Essl, "Cross-corpus Acoustic Emotion Recognition with Multi-task Learning: Seeking Common Ground while Preserving Differences," *IEEE Transactions on Affective Computing*, no. 99, 2017, 14 pages.

[3] M. Najafabadi, F. Villanustre, T. M. Khoshgoftaar, N. Seliya, R. Wald, and E. Muharemagic, "Deep learning applications and challenges in big data analytics," *Journal of Big Data*, vol. 2, no. 1, p. 1, February 2015.

[4] "Hours of video uploaded to YouTube every minute as of July 2015," https://www.statista.com/topics/2019/youtube, accessed October 5, 2016.

[5] A. Burmania, M. Abdelwahab, and C. Busso, "Tradeoff between quality and quantity of emotional annotations to characterize expressive behaviors," in *Proc. of ICASSP*, March 2016, pp. 5190–5194.

[6] Z. Zhang, N. Cummins, and B. Schuller, "Advanced Data Exploitation in Speech Analysis – An Overview," *IEEE Signal Processing Magazine*, vol. 34, July 2017, 24 pages.

[7] W. Han, E. Coutinho, H. Ruan, H. Li, B. Schuller, X. Yu, and X. Zhu, "Semi-Supervised Active Learning for Sound Classification in Hybrid Learning Environments," *PLoS ONE*, vol. 11, no. 9, 2016.

[8] S. Hantke, F. Eyben, T. Appel, and B. Schuller, "iHEARu-PLAY: Introducing a game for crowdsourced data collection for affective computing," in *Proc. of the Int. Workshop on Automatic Sentiment Analysis in the Wild held in conjunction with the biannual Conf. on Affective Computing and Intelligent Interaction*. Xi'an, P. R. China: IEEE, September 2015, pp. 891–897.

[9] P. Hsueh, P. Melville, and V. Sindhwani, "Data quality from crowdsourcing: A study of annotation selection criteria," in *Proc. of the NAACL HLT Workshop on Active Learning for Natural Language Processing*. Boulder, USA: ACL, June 2009, pp. 27–35.

[10] A. Burmania, S. Parthasarathy, and C. Busso, "Increasing the reliability of crowdsourcing evaluations using online quality assessment," *IEEE Transactions on Affective Computing*, vol. 7, no. 4, pp. 374–388, Oct 2016.

[11] R. Morris, D. McDuff, and R. Calvo, "Crowdsourcing techniques for affective computing," in *Handbook of Affective Computing*, ser. Oxford Library of Psychology, R. A. Calvo, S. D'Mello, J. Gratch, and A. Kappas, Eds. Oxford University Press, 2015, pp. 384–394.

[12] A. Tarasov, S. J. Delany, and C. Cullen, "Using crowdsourcing for labelling emotional speech assets," in *Proc. W3C workshop on Emotion Markup Language (EmotionML)*. Paris, France: Springer, 2010, pp. 1–5.

[13] P. Covington, J. Adams, and E. Sargin, "Deep neural networks for youtube recommendations," in *Proc. of the ACM Conf. on Recommender Systems*. New York, USA: ACM, September 2016, pp. 191–198.

[14] D. Watts and S. Strogatz, "Collective dynamics of small-world networks," *Nature*, vol. 393, no. 6684, pp. 440–442, 1998.

[15] S. Strogatz, "Exploring complex networks," *Nature*, vol. 410, no. 6825, pp. 268–276, 2001.

[16] M. Newman, D. Watts, and S. Strogatz, "Random graph models of social networks," *Proceedings of the National Academy of Sciences*, vol. 99, no. suppl 1, pp. 2566–2572, 2002.

[17] A. Härmä, M. F. McKinney, and J. Skowronek, "Automatic surveillance of the acoustic activity in our living environment," in *Proc. of the Int. Conf. on Multimedia and Expo*. Amsterdam, The Netherlands: IEEE, July 2005, no pagination.

[18] B. Schuller, S. Steidl, and A. Batliner, "The Interspeech 2009 Emotion Challenge," in *Proc. of INTERSPEECH*. Brighton, UK: ISCA, September 2009, pp. 312–315.

[19] M. Schmitt and B. W. Schuller, "openXBOW — Introducing the Passau open-source crossmodal bag-of-words toolkit," *Journal of Machine Learning Research*, 2017, 5 pages.

[20] Y. LeCun, B. Boser, J. Denker, D. Henderson, R. Howard, W. Hubbard, and L. Jackel, "Backpropagation applied to handwritten zip code recognition," *Neural computation*, vol. 1, no. 4, pp. 541–551, 1989.

[21] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proc. of the 32nd Int. Conf. on Machine Learning*. Lille, France: ACM, 2015, pp. 448–456.

[22] G. Keren, J. Deng, J. Pohjalainen, and B. Schuller, "Convolutional neural networks with data augmentation for classifying speakers native language," in *Proc. of INTERSPEECH*. San Francisco, CA, USA: ISCA, September 2016, pp. 2393–2397.

[23] S. Amiriparian, M. Gerczuk, S. Ottl, N. Cummins, M. Freitag, S. Pugachevskiy, and B. Schuller, "Snore Sound Classification Using Image-based Deep Spectrum Features," in *Procc. of INTERSPEECH*. Stockholm, Sweden: ISCA, August 2017, 5 pages.

[24] N. Cummins, S. Amiriparian, G. Hagerer, A. Batliner, S. Steidl, and B. Schuller, "An Image-based Deep Spectrum Feature Representation for the Recognition of Emotional Speech," in *Proc. of the 25th ACM International Conference on Multimedia, MM 2017*. Mountain View, CA: ACM, October 2017, 7 pages.