INTELLIGENT SIGNAL PROCESSING MECHANISMS FOR NUANCED ANOMALY DETECTION IN ACTION AUDIO-VISUAL DATA STREAMS

Josef Kittler[*], *Ioannis Kaloskampis*[†], *Cemre Zor*[*], *Yong Xu*[*], *Yulia Hicks*[†] *and Wenwu Wang*[*]

[*]Centre for Vision, Speech and Signal Processing, University of Surrey, UK
[†]Cardiff School of Engineering, Cardiff University, UK
{J.Kittler, C.Zor, Yong.Xu, W.Wang}@surrey.ac.uk, {KaloskampisI, HicksYA}@cardiff.ac.uk

ABSTRACT

We consider the problem of anomaly detection in an audio-visual analysis system designed to interpret sequences of actions from visual and audio cues. The scene activity recognition is based on a generative framework, with a high-level inference model for contextual recognition of sequences of actions. The system is endowed with anomaly detection mechanisms, which facilitate differentiation of various types of anomalies. This is accomplished using intelligence provided by a classifier incongruence detector, classifier confidence module and data quality assessment system, in addition to the classical outlier detection module. The paper focuses on one of the mechanisms, the classifier incongruence detector, the purpose of which is to flag situations when the video and audio modalities disagree in action interpretation. We demonstrate the merit of using the Delta divergence measure for this purpose. We show that this measure significantly enhances the incongruence detection rate in the Human Action Manipulation complex activity recognition data set.

Index Terms— Audio visual scene analysis, incongruence detection, anomaly detection

1. INTRODUCTION

The problem of anomaly detection in signal processing has recently been receiving increasing attention. An anomaly is defined in the literature as an outlier from some known distribution [1, 2]. It typically raises some cause for concern, and could be indicative of anomalous events such as an engine failure, a medical problem, or a cyber attack.

An important medium for anomaly detection is video. It is currently one of the fastest growing data resources, mainly due to the wide use of video surveillance and the extensive growth of video content on the web. This growth has created the need for automatic detection of anomalous events in the video content, which could translate into a burglary, a terrorist attack, a fight, a mistake during the execution of an activity or inappropriate content which needs to be flagged for removal.

The audio cue is inextricably linked to the events taking place in the scene. An object in the scene produces distinctive sounds when it is interacted with, which are related to the object's material or the actions which caused the impact [3]. Audio can offer vital information in the scene, especially when the object of interest is occluded or even outside the camera's field of view. Therefore, its contribution towards the recognition of unusual events can be very important. Nevertheless, there is little work in the literature that investigates the problem of anomaly detection with the combined use of video and audio information.

The reason behind the lack of systems which consider both the audio and the video cue for anomaly detection could potentially lie in the absence of tools which enable the analysis of the decision making processes in multimodal systems. A recently developed mechanism which is relevant to the aforementioned task is classifier incongruence detection, which gauges the consistency of classifier outputs [4]. Normally all classifiers analysing a scene should support a specific decision. Therefore, incongruence of classifiers could be indicative of an anomaly. Incongruence between different modalities could translate into a sensor malfunction or a spoofing attempt.

In this paper, we develop a system for automatically detecting incongruous outputs of multimodal classifiers interpreting audio-visual scenes in the context of complex human activity recognition. Towards this effort, we build contextual classifiers which recognise human actions based on audio and video features and detect incongruence between different modalities with the recently proposed Delta divergence [4]. We show that its performance is superior to the classical Kullback-Leibler (KL) divergence [5] and its decision cognizant variant DC-KL [6].

The rest of our article is structured as follows. First, we discuss research related to our work in Section 2. The proposed system for detecting incongruence of classifiers interpreting audio-visual data streams is described in Section 3. The experiments conducted on the publicly available Human

This work was supported by the Engineering and Physical Sciences Research Council (EPSRC) Grant number EP/K014307 and the MOD University Defence Research Collaboration in Signal Processing.

Action Manipulation complex activity recognition data set [7] are presented in Section 4. We conclude in Section 5.

2. RELATED WORK

Our work is related to the research area of complex human behaviour analysis based on audio and video cues. Interestingly, most of the work carried out in this domain focuses on monitoring cooking activities. Pieropan et al. [7] investigated the problem of recognising simple actions, such as *pour milk, open cereal box* occurring during the complex activity of preparing cereals. They used RGB-D video as well as audio and utilised a hidden Markov model (HMM) to model the complex activity. Kojima et al. [8] employed a framework consisting of convolutional neural networks and hierarchical HMMs which recognised cooking recipes and could operate as a cooking support system. Salvi et al. [9] used a discrete Bayesian network to teach a robot to map spoken words to the execution of actions and various environment entities.

Even though our system integrates the processes of action and activity recognition, the focus of our work is the detection of anomalies. Here, we review certain methods that are close to our research. For a more comprehensive coverage of the area we refer the reader to [10, 11, 12].

Our system detects multimodal classifier incongruence as a prerequisite to anomaly detection. The key statistic for this purpose is divergence, that is, a measure of difference between two aposteriori class probability distributions. A prominent measure of divergence is the Kullback-Leibler (KL) divergence [13]. KL is widely used for incongruence detection between two classifiers [14] and Itti and Baldi refer to it with the term Bayesian surprise [5]. The main disadvantage of the KL divergence is that it gives all class probabilities equal consideration without distinguishing between dominant and non-dominant hypotheses. This treatment may give inappropriate weight to clutter in multi-class problems. To avoid this problem, Ponti et al. introduced the decision cognizant variant of KL (DC-KL) [6]. The idea behind DC-KL is to merge all non-dominant hypotheses into a single hypothesis in an attempt to reduce the amount of clutter. However, DC-KL inherits some properties from KL which limit its robustness in detecting incongruence. Specifically, (i) the measures defined by taking different distributions as a reference are non-symmetric, (ii) the divergence function may give the same value for different distributions and therefore it does not distinguish between congruence and incongruence unambiguously, (iii) it involves the ratio of a posteriori probabilities of the two classifiers which may approach infinity for zero denominator and dominate the final value of the measure. To overcome these issues with the KL and DC-KL measures, Kittler and Zor [4] proposed Delta divergence, a decision cognizant measure of classifier incongruence based on quadratic rather than logarithmic entropy. The sensitivity of Delta divergence to estimation errors was studied in [15].

Classifier incongruence detection was applied to the problem of detecting instances of a novel class by monitoring the outputs of generic and specific object category classifiers [16]. Detecting incongruence between contextual and non-contextual classifiers was shown to play an important role in automatic tennis video interpretation in [12] and activity detection in video in [15]. Herein we study the problem of detecting incongruence between multimodal classifiers in the context of scene understanding of complex human behaviour from audio and video streams. The aim is to use the detected incongruences to trigger further investigative mechanisms in order to pinpoint the nature of anomalies which gave rise to the disparate classifier opinions. These would include data quality detection, flagging issues such as sensor failure, unfavourable environmental conditions, anomalous events occurring in the scene, etc. For detecting incongruences, we primarily rely on the Delta divergence and show its superior ability to detect true incongruences.

3. METHODOLOGY

Our system recognises complex human activities in audio-visual scenes from multimodal observations. It is endowed with the ability to detect incongruence between the modalities to flag potential anomalies. A classifier for each modality is built using the respective audio and video training data. We utilise the recently introduced Delta divergence to measure classifier incongruence and relate it to a threshold guaranteeing a specified level of confidence in not rejecting true congruences.

The concept of complex behaviour implies that each activity performed by a human consists of several steps, which we call actions [17]. For example, the complex activity of *high jump* consists of the actions *running, jumping and falling*. To detect an activity's constituent actions we need to determine the temporal boundaries for each action, *i.e.* its start and end points within the video or audio stream.

Our system works as follows. First, low level features are extracted from each data stream. These features are then converted into mid-level representations which discover underlying patterns within the data and ease the classification task. A widely adopted mid-level representation, which we use in this work, is the Fisher representation [18]. Using the mid-level representation we build a contextual classifier for each stream. The contextual classifiers utilise context free grammars and HMMs to detect actions within the input streams and specify their temporal boundaries. Therefore, our system converts an input stream to temporal segments and assigns each temporal segment to an action class.

The output of the contextual classifiers is then utilised to detect incongruence as follows. An action segment is represented by an observation vector $\mathbf{x}$ belonging to one of mutually exclusive action classes $\omega_i, i = 1, ..., m$. The observation vector $\mathbf{x}$ is the segment's visual feature representation.

Given observation $\mathbf{x}$, we denote the aposteriori probability of its membership in class ω_i by $P(\omega_i|\mathbf{x})$. The vector $\mathbf{x}$ is automatically assigned to one of the action classes by the video classifier. We assume that the classifier effectively computes the aposteriori class probabilities $P(\omega_i|\mathbf{x}), \forall i$ and engages a Bayesian decision rule to effect the class assignment.

The audio classifier makes a decision regarding action segment's class based on its set of aposteriori class probabilities $\tilde{P}(\omega_i|\mathbf{y}), \forall i$, based on observation $\mathbf{y}$. The observation vector $\mathbf{y}$ is the segment's audio feature representation. We would like to measure the incongruence of the video and audio classifiers given the observations $\mathbf{x}$ and $\mathbf{y}$. Given the two probability distributions, we would consider the classifiers congruent if the two probability distributions agree, and incongruent, in the case that the two probability distributions disagree. For simplicity, we will no longer refer to the observations $\mathbf{x}, \mathbf{y}$ explicitly and denote the class probabilities by P_i and $\tilde{P}_i$, so that:

$$\begin{aligned} P_i &= P(\omega_i|\mathbf{x}) \\ \tilde{P}_i &= \tilde{P}(\omega_i|\mathbf{y}), \forall i \end{aligned} \tag{1}$$

When comparing the outputs of the video and audio classifiers, only three outcomes interest us: the dominant class ω identified by the classifier with probability distribution P, the dominant class $\tilde{\omega}$ identified by the other classifier, and neither of the two, in other words $\hat{\omega} = \Omega - \omega - \tilde{\omega}$. Based on this notion, the Delta divergence, D_Δ is defined as:

$$D_\Delta = \frac{1}{2}\left[\sum_{i\in\{\omega,\tilde{\omega}\}} |\tilde{P}_i - P_i| + |\tilde{P}_{\hat{\omega}} - P_{\hat{\omega}}|\right] \tag{2}$$

where $P_{\hat{\omega}} = 1 - P_\omega - P_{\tilde{\omega}}$. We compare this measure with the conventional Kullback-Leibler measure:

$$D_{KL} = \sum_i \tilde{P}_i \log \frac{\tilde{P}_i}{P_i} \tag{3}$$

and its decision cognizant variant:

$$D_{DCKL} = \sum_{i\in\{\omega,\tilde{\omega}\}} \tilde{P}_i \log \frac{\tilde{P}_i}{P_i} + \tilde{P}_{\hat{\omega}} \log \frac{\tilde{P}_{\hat{\omega}}}{P_{\hat{\omega}}} \tag{4}$$

4. ACTION RECOGNITION EXPERIMENT

We evaluate our framework on the Human Action Manipulation dataset compiled to investigate the problem of action recognition from video and audio streams. In this dataset eight subjects carry out the task of preparing cereals. Each execution is recorded in a video and an audio stream. The participants are not instructed on how to execute the task and consequently the task is performed in several different ways. There are six actions that are generally followed, *i.e. open milk box, pour milk, close milk box, open cereal box, pour cereals and close cereal box.* There is a variability in terms of the order these actions are carried out by each participant. In a number of executions some of these actions are omitted. The goal is to detect and recognise all actions occurring in the video and audio streams.

To detect the actions in the video, we work in similar fashion to [19]. We first extract low-level local features with improved dense trajectories (iDTFs) [20]. The dimensions of iDTFs are reduced from 426 to 213 with PCA. The reduced size features are then converted to Fisher vectors (FVs) [18] as follows: first, the training and test subsets are determined; then, 260000 features are selected at random from the training subset and they are clustered into 16 clusters with the GMM algorithm; using these clusters, all reduced-size features are encoded to FVs with the VLFeat toolbox [21]; finally, L2 and power normalisations are applied to the FVs. The resulting FVs are of size $2 * K * D$, where K is the number of clusters of the GMM and D the dimensions of the reduced size iDTF descriptor. In our case, for $K = 16$ and $D = 213$ the size of each FV is 6816 dimensions, which we reduce to 64 dimensions with a second PCA. Having obtained the reduced FVs, we recognise actions in the video stream with the HTK toolkit [22]. To detect the actions in audio, we first extract Mel-Filter banks (MFB) [23] with 40 channels from the 16kHz sample rate waveforms. The short-time Fourier transform (STFT) is first applied to obtain the spectral features. Then, the Mel-filters are applied on the magnitude spectrogram to get the final audio features. Having obtained the audio features, actions in the audio stream are recognised with the HTK toolkit. For each stream and each detected action, the HTK toolkit provides its temporal extend (*i.e.* its start and end point within the video or audio), its class (*e.g. close milk box, open cereal box*) and a detection score in the form of log-likelihood.

We measure the performance of the video and audio classifiers in terms of per frame accuracy. Following the recommendations from [7], we estimate the accuracy for ten random splits of the dataset into training and testing subsets and average out the results to obtain the total classification accuracy. Our video classifier achieves an accuracy of 91.5% while the accuracy of the audio classifier is 65.4%. The state-of-the-art performance for this dataset is 73% [7].

For the incongruence detection experiments we do not take random splits. Rather, we split the dataset five times into training and test subsets so that each video appears once in the testing subset. The average per frame classification accuracy for these five splits is 90.16% and 69.92% for the video and audio classifier respectively.

4.1. Incongruence detection results

The multimodal audio video interpretation system described above has been augmented by an incongruence detector to monitor the outputs of the two monomodal classifiers. Three implementations have been experimented with, realising the

	95% Conf.	90% Conf.
# Incongruences by Δ_D	69	95
# Incongruences by $DCKL$	34	57
# Incongruences by KL	14	16

Table 1. Number of correctly identified incongruences (true negatives) by Δ_D, $DCKL$ and KL for 90% and 95% confidence intervals, out of 183 incongruent cases.

incongruence measures (2), (3) and (4) respectively. In order to flag incongruences a decision threshold has to be defined. Note that each incongruence measure is a statistic with a certain distribution. Ideally the threshold should be set so as to recognise all congruent outputs as such. However, in practice the distribution will have tails and, as a compromise, we wish to set the threshold so that the majority of congruent cases are accepted by the detector at a given level of confidence. As in the case of outlier detection, the threshold is set at a user specified level of confidence.

We have experimented with a range of confidence level values. The corresponding classifier incongruence rates are given in Table 4.1. The table shows the detection rates achieved with the three different measures tested on 183 incongruent cases. These are defined as all the cases which were assigned different class identities by the two classifiers. From the results it is evident that the the Delta divergence is able to detect almost twice as many true incongruences. The corresponding ROC curve is shown in Fig. 1 where the false positive (false alarm) rates are calculated for a variety of confidence levels (true negative rates) within the range [0, 1].

The selected thresholds miss some incongruences and these will become false negatives. This occurs because the audio data is somewhat ambiguous and often the audio classifier misclassifies some actions. This would not necessarily be a problem. However, the audio classifier tends to drive the posterior class probability of the winning hypothesis close to one, i.e., the classifier is overconfident, even when it is wrong. Therefore, the performance of the incongruence detector improves when the classifiers provide a realistic assessment of their competences. Our future work will investigate deep neural network alternatives to improve the system performance.

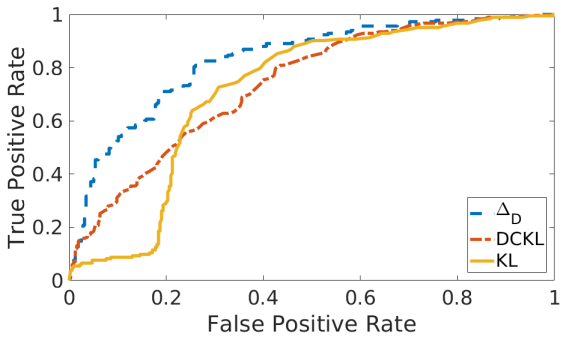

Fig. 1. ROC analysis for Δ_D, $DCKL$ and KL

Fig. 2. Key frame for incongruent classifier decisions

Once a case of incongruence is detected, it gives the operator the possibility to take an appropriate action. For example, consider the video segment represented by the key frame shown in Fig. 2. Here, the audio modality worked well and the cause for incongruence was the out of vocabulary event involving multiple cereal boxes, which the video system could not interpret correctly, as it was not trained for such eventuality. The outcome of the incongruence detection and analysis would be to retrain the visual model with samples which illustrate multiple instances of the same object class in the scene.

5. CONCLUSION

We investigated the problem of classifier incongruence detection in the context of a multimodal human action recognition system deployed in a kitchen activity interpretation scenario, as a mechanism facilitating comprehensive anomaly detection. Classifier incongruence applied to audio and video modalities can indicate a sensor failure, a change in environmental conditions, out of vocabulary scene content, occlusion and other types of anomalies. We used the recently proposed Delta divergence as a classifier incongruence measure and demonstrated its superior ability to detect true incongruences at any specified level of confidence, in comparison to the conventional Kullback-Leibler measure and its decision cognizant variant. Our approach flagged interesting anomalies which may be of interest in routine operation, or which question the underlying models of the scene interpretation system. The latter is exemplified by the presence of unexpected scene objects, for which the interpretation system should develop appropriate models. One of the challenges of the proposed approach is the assumption that at least one classifier identifies the true action correctly. When both classifiers make an error, resulting in their outputs being congruent, the classifier failures will not be detected. In future we shall explore complementary mechanisms, such as classifier confidence assessment and data quality assessment, to establish, whether they would offer a comprehensive solution in such cases.

6. REFERENCES

[1] D. Agarwal, "Detecting anomalies in cross-classified streams: a bayesian approach," *Knowledge and Information Systems*, vol. 11, no. 1, pp. 29–44, Jan 2007.

[2] F. J. Anscombe and Irwin Guttman, "Rejection of outliers," *Technometrics*, vol. 2, no. 2, pp. 123–147, 1960.

[3] A. Owens, P. Isola, J. McDermott, A. Torralba, E. H. Adelson, and W. T. Freeman, "Visually indicated sounds," in *Proc. IEEE CVPR*, June 2016, pp. 2405–2413.

[4] J. Kittler and C. Zor, "Delta divergence: A novel decision cognizant measure of classifier incongruence," 2016, arXiv:1604.04451.

[5] L. Itti and P. F. Baldi, "A principled approach to detecting surprising events in video," in *Proc. IEEE CVPR*, 2005, pp. 631–637.

[6] M. Ponti, J. Kittler, M. Riva, T. de Campos, and C. Zor, "A decision cognizant Kullback-Leibler divergence," *Pattern Recognition*, vol. 61, pp. 470–478, 2017.

[7] A. Pieropan, G. Salvi, K. Pauwels, and H. Kjellström, "Audio-visual classification and detection of human manipulation actions," in *Proc. IEEE/RSJ IROS*, Sept 2014, pp. 3045–3052.

[8] R. Kojima, O. Sugiyama, and K. Nakadai, "Audio-visual scene understanding utilizing text information for a cooking support robot," in *Proc. IEEE/RSJ IROS*, Sept 2015, pp. 4210–4215.

[9] G. Salvi, L. Montesano, A. Bernardino, and J. Santos-Victor, "Language bootstrapping: Learning word meanings from perception x2013;action association," *IEEE Trans. SMC, Part B (Cybernetics)*, vol. 42, no. 3, pp. 660–671, June 2012.

[10] V. Chandola, A. Banerjee, and V. Kumar, "Anomaly detection: A survey," *ACM Comput. Surv.*, vol. 41, no. 3, pp. 15:1–15:58, July 2009.

[11] S. Agrawal and J. Agrawal, "Survey on anomaly detection using data mining techniques," *Procedia Computer Science*, vol. 60, no. Supplement C, pp. 708 – 713, 2015.

[12] J. Kittler, W. Christmas, T. de Campos, D. Windridge, F. Yan, J. Illingworth, and M. Osman, "Domain anomaly detection in machine perception: A system architecture and taxonomy," *IEEE Trans. PAMI*, vol. 36, no. 5, pp. 845–859, May 2014.

[13] S. Kullback and R. A. Leibler, "On information and sufficiency," *The Annals of Mathematical Statistics*, vol. 22, no. 1, pp. 79–86, 1951.

[14] D. Weinshall, A. Zweig, H. Hermansky, S. Kombrink, F. W. Ohl, J. Anemller, J. H. Bach, L. Van Gool, F. Nater, T. Pajdla, M. Havlena, and M. Pavel, "Beyond novelty detection: Incongruent events, when general and specific classifiers disagree," *IEEE Trans. on PAMI*, vol. 34, no. 10, pp. 1886–1901, Oct 2012.

[15] J. Kittler, C. Zor, I. Kaloskampis, Y. Hicks, and W. Wang, "Error sensitivity analysis of delta divergence - a novel measure for classifier incongruence detection," *Pattern Recognition*, vol. 77, pp. 30 – 44, 2018.

[16] D. Coppi, T. de Campos, F. Yan, J. Kittler, and R. Cucchiara, "On detection of novel categories and subcategories of images using incongruence," in *Proc. ICMR*, April 2014, p. 337.

[17] I. Kaloskampis, Y. Hicks, and D. Marshall, "Automatic analysis of composite activities in video sequences using key action discovery and hierarchical graphical models," in *Proc. IEEE ICCV Workshops*, Barcelona, Spain, Nov. 2011, pp. 890– 897.

[18] F. Perronnin, J. Sánchez, and T. Mensink, "Improving the fisher kernel for large-scale image classification," in *Proc. ECCV*, 2010, pp. 143–156.

[19] H. Kuehne, J. Gall, and T. Serre, "An end-to-end generative framework for video segmentation and recognition," in *Proc. IEEE WACV*, Lake Placid, Mar 2016.

[20] H. Wang and C. Schmid, "Action recognition with improved trajectories," in *2013 IEEE International Conference on Computer Vision*, Dec 2013, pp. 3551–3558.

[21] A. Vedaldi and B. Fulkerson, "VLFeat: An open and portable library of computer vision algorithms," `http://www.vlfeat.org/`, 2008.

[22] Machine Intelligence Laboratory of the Cambridge University Engineering Department, "The hidden Markov model toolkit (HTK)," `http://htk.eng.cam.ac.uk/`, 2016.

[23] S. K. Kopparapu and M. Laxminarayana, "Choice of Mel filter bank in computing MFCC of a resampled speech," in *Proc. ISSPA*, May 2010, pp. 121–124.