DCU team at the NTCIR-15 Micro-Activity Retrieval Task

Tu-Khiem Le*
The Insight Centre for Data Analytics,
Dublin City University
Dublin, Ireland
tu-khiem.le4@mail.dcu.ie

Manh-Duy Nguyen*
Dublin City University
Dublin, Ireland
manh.nguyen5@mail.dcu.ie

Ly-Duyen Tran
Dublin City University
Dublin, Ireland
ly.tran2@mail.dcu.ie

Van-Tu Ninh
Dublin City University
Dublin, Ireland
van.ninh2@mail.dcu.ie

Cathal Gurrin
Dublin City University
Dublin, Ireland
cathal.gurrin@dcu.ie

Graham Healy
Dublin City University
Dublin, Ireland
graham.healy@dcu.ie

ABSTRACT

The growing attention to lifelogging research has led to the creation of many retrieval systems, most of which employed event segmentation as a core functionality. While previous literature focused on splitting lifelog data into broad segments of daily living activities, less attention was paid to micro-activities which last for short periods of time, yet carry valuable information for building a high-precision retrieval engine. In this paper, we present our efforts in addressing the NTCIR-15 MART challenge, in which the participants were asked to retrieve micro-activities from a multi-modal dataset. We proposed five models which investigate imagery and sensory data, both jointly and separately using various Deep Learning and Machine Learning techniques, and achieved a maximum mAP score of 0.901 using an Image Tabular Pair-wise Similarity model, and overall ranked second in the competition. Our model not only captures the information coming from the temporal visual data combined with sensor signal, but also works as a Siamese network to discriminate micro-activities.

TEAM NAME

DCU

SUBTASKS

Retrieval task

1 INTRODUCTION

The continuous development of technology has led to an increase in the availability of many low-cost personal devices, which aid people in digitalising their life, either actively or passively. One can now effortlessly construct detailed digital records by aggregating data sources from various wearable devices, where such a person is known as a lifelogger [14]. Lifelog data is a type of multi-modal big data that consists of spatio-temporal information (e.g. timezone, geographic location), visual information (e.g. images, videos) and bio-metric states of the body (e.g. heart-rate, calorie intake, steps). Owing to the huge volume of data generated, there has been an increasing need for efficient data management and effective retrieval systems. Consequently, many challenges were held to tackle this problem, namely ImageCLEFlifelog [3–5, 28], NTCIR Lifelog Task [8–10] and Lifelog Search Challenge (LSC) [11, 12], each of which attracted many researchers to develop search engines, and to evaluate these retrieval systems based on various metrics of performance.

The recognition of daily human activities plays an important role in many lifelog retrieval systems [14]. Such activities are employed to segment data, but only into large chunks such as working, driving, or eating. In fact, each chunk actually comprises many smaller activities which carry more detail about the lifeloggers' activities, and this contributes directly to the accuracy of the retrieval systems. For example, one might be replying to emails, organizing documents or answering a phone within an activity segment labelled as "working". With the availability of lifelog data in the past work, a retrieval system would not indicate precisely which period of the activity corresponds to "replying email" activity. Therefore, there is a need for an efficient multi-modal source of lifelog data which can capture and distinguish the totality of human activities that can occur in daily life.

The NTCIR Micro-Activity Retrieval Task (MART) [17] focused on the identification of micro activities, which only happen within a short time period are not typically identified using conventional wearable devices. The task organisers have introduced the first generation of a new micro-activities in daily living dataset which incorporates various data sources: images from lifelogging camera, electrooculogram (EOG) signals from eye movements, heart rate, accelerometer readings of two hands and head movements, and computer interaction data from Loggerman software [1]. The combination of these sources provided rich insight into the action being performed by the lifelogger. For instance, the hands' movement (accelerometers) can differentiate the activity of typing on a keyboard and answering a phone call, or the movement of eyes (EOG) can reveal signatures related to different types of activities (e.g. reading).

In the MART retrieval task, participants were asked to predict 20 classes of micro-activity performed by 7 different subjects, either automatically or interactively. As there are 140 items to be predicted in the test set, the task participants needed to submit 20 ranked lists for each submission, each of which contained 140 ids in the order of relevance, making a total of 2,800 rows in the submission file. In this paper, we present five approaches to addressing the MART retrieval challenge. Two of our approaches, BI (Baseline Image) model and BT (Baseline Tabular) model exploit image and tabular

*Both authors contributed equally to this paper

[1] http://loggerman.org/

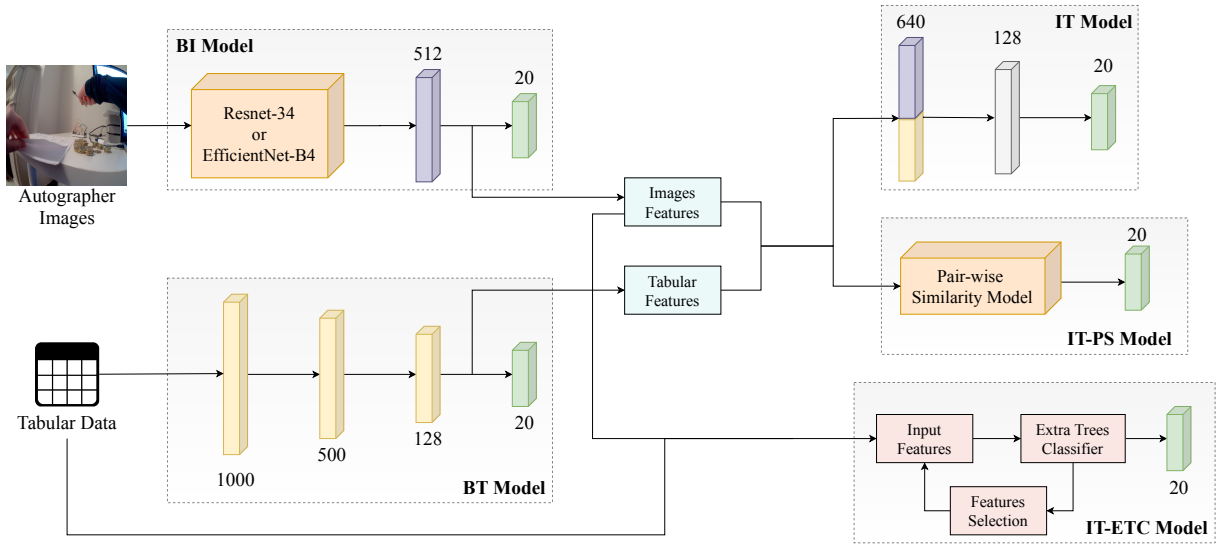

Figure 1: Overview of five proposed models in a complete pipeline

data separately, the other three approaches examine different ways of combining these two approaches to predict the corresponding micro-activity. A post-processing procedure was applied for all proposed models to obtain an optimised ranked list for submission.

2 RELATED RESEARCH

With the development of a wide range of sensors for human activity, much research has been conducted to recognise basic human activities using these; for example using accelerometers, gyroscopes and magnetometers [38]. With the popularity of smartphones, many other diverse data sources can be captured using sensors like proximity sensors, ambient light sensors, thermometer and barometer sensors. All of these sensors can be employed to build sensor fusion-based human activity recognition models with higher accuracy [31] and more energy-efficient [32] than using only one single data source such as from an accelerometer [22, 24, 26, 33]. Much research has also been done to show that Machine Learning models work effectively on these problems [1, 21, 35], however, Hassan et al. suggest that Deep Learning can be more robust that traditional typical approaches such as Support Vector Machine (SVM) and Artificial Neural Network (ANN) [15]. The same work which utilises Convolutional Neural Network (CNN) proposed by Ignatov on two commonly used WISDM and UCI datasets also achieve state-of-the-art performance while requiring low computational cost and no manual feature engineering [19]. Furthermore, Wang et al. suggest that hybrid Deep Learning models, which combine CNN and Recurrent Neural Networks (RNN) [29, 34, 39], have the tendency to perform better than single Deep Learning models [38].

Besides the use of sensors data for activity recognition, electrooculography (EOG) is also considered to be a potential sensor source from which to identify which actions are being performed by human when eye movement behaviour contains discriminative information. Andreas Bulling et al. [2] was one of the pioneers in exploiting eye movement for activity recognition. The authors succeeded to classify 5 types of activity (copying text, reading paper, taking handwritten notes, watching a video, browsing the web) by examining different eye movement features extracted from EOG data (such as blinks, saccades, fixations). Shoya Ishimaru et al. [20] also achieved compelling results by using portable commercial EOG glasses to classify 4 activities (eating, talking, typing, reading). These prior works demonstrate that the eye movement patterns can reveal important properties and behaviours of certain actions, which can lead to an accurate prediction of human activities by combining them with other sensors' measurements.

Similar with MART, the LSC (Lifelog Search Challenge) is also a retrieval benchmarking activity for multi-modal lifelog data but mostly relies on visual, geospatial and biometric data [13]. Unlike MART, participants in LSC, however, need to build interactive systems to assist a user to find a specific activity matching a given textual description. Many such interactive systems have the option for users to query by sketching an overview of an activity [18] while virtual environment also showed its own advantages [7]. Ranking algorithms for documents retrieval can also be applied [37]. Converting the query and visual data into the same space to be able to make the comparison is also another point of view [25]. Besides, the system of Mejzlík et al. [27] supports the option for retrieving based on user's relevance feedback. Many of above systems are based on the indexing of visual concepts detected from image data with other sources of evidence used as filters, and serve as interactive interrogation tools for individuals to use. Our approach, in contrast, is entirely automatic and also employs additional sensor signal information. Moreover, we do not utilise the object detection information in the MART metadata but use entire images as one of our input features.

Our main contributions in this paper are threefold. Firstly, we introduce two simple approaches, which are considered baseline approaches, which utilise only either visual or sensory data. Secondly, we integrate both of these types of data to create three improved models that use both deep neural networks and a traditional tree

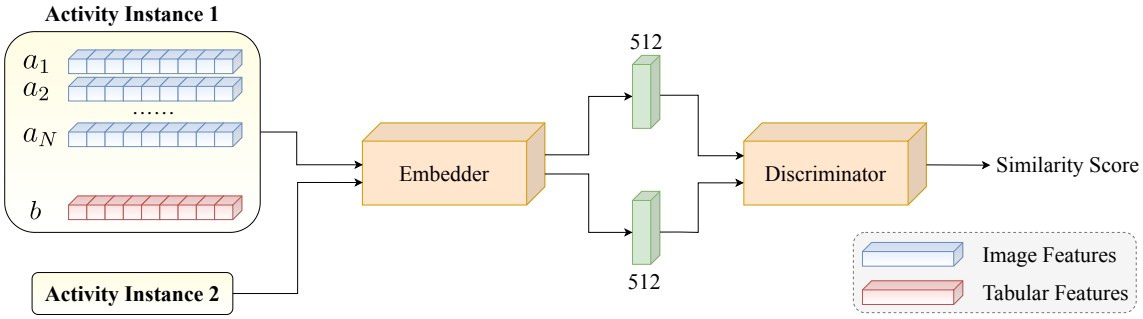

Figure 2: Image Tabular Pair-wise Similarity Model Architecture

classifier with a feature selection process. Thirdly, our submission runs shed light on the suggestion that including non-visual sensory data alongside visual data, is advantageous for MART in which our best submission run's model achieved second place with the score of 0.901.

3 METHOD

3.1 Overview

The provided MART dataset consists of four main parts which are: images from an Autographer (lifelog camera), tabular data of pre-computed features, screenshots and raw data (time-series data from recording devices for who do not wish to use pre-computed features). Among these, we have a particular interest in the images from the Autographer and the tabular data of sensor reading features. Images are visually rich as they contain valuable information such as objects appearing therein and scene changes, which might offer clues about the activity's type. Tabular data, on the other hand, reflects the internal properties of each activity through the interaction between the subjects and the environments such as how they moved their hands or where they looked at. In contrast, we do not consider screenshots to be of benefit to the retrieval task as they are likely to yield biases and noises in the training process. This is because screenshots in the dataset were being captured continuously even when the subjects don't interact with the computer.

We aimed at developing simple classification models for image and tabular data to demonstrate the possibility of utilising these features to infer human activities. Upon this, we further investigated more advanced methods which aggregate image and tabular features in several ways in order to optimise the classification results. Figure 1 provides an overview of the structure of our proposed models.

3.2 Baseline Image Model (BI)

In this model, we proposed a simple Image classification model using only the data from the Autographer as input. Since an event is depicted by a sequence of images in chronological order, we decided to consider the images independently to perform the classification and average the score across all images to obtain the prediction afterwards.

Regarding the size of the input, which is relatively small (2,752 images), it is advantageous to apply a small-scale network in order to facilitate a fast training, small memory, yet good performance model. Resnet-34 [16] is on top of our consideration as deeper models tend to be overfitting easily given the small data size. Moreover, we also compare Resnet-34 with EfficientNet-B4 [36] model which has slightly less parameters (around 19 millions) but was reported to achieve higher classification result. The BI model was created by using those models as a backbone and adding a hidden fully connected layer after the global pooling layer of them. The output of this hidden layer would work as a features for later advanced methods.

3.3 Baseline Tabular Model (BT)

Similarly, we developed a Tabular model in which sensors and computer interaction data utilised as input. The task organiser has provided a CSV file containing some basic statistics (min, max, median, average, standard deviation) extracted from raw signal data, which are handy and easy to use features for micro-activity prediction. Since we only interest in data from sensors and Loggerman (computer interaction tracker), the majority part of the CSV file regarding the Resnet-101 features was eliminated.

We pre-processed the given data by filling the missing records by the median value and creating a corresponding columns with True/False value to keep track of the filled positions. The data was then normalised before feeding to the Tabular model. The model is a simple neural network with the use of three hidden layers in which the output of the last layer would be a tabular features for our advanced architectures.

3.4 End-to-end Image Tabular Model (IT)

Upon two proposed baseline approaches, we developed a hypothesis that the combination of both features (images and sensors) would yield a better result. To verify our assumption, we advanced the model by concatenating the Images features and Tabular features produced by BI and BT model. They were followed by a hidden layer and performed end-to-end training to predict the micro-activities. The detailed architecture of this model is illustrated in Figure 1. It was noted that the IT model included both BI and BT baseline, hence the weights of this model were initialised from two previous trained models and all would be updated in the training stage.

Table 1: Micro-Activity Retrieval Results

Model	Input	Structure	mAP
BT	Tabular data	MLP	0.69817
BI	Image data	Resnet-34	0.70279
		EfficientNet-B4	0.73694
IT	Tabular and Image data	MLP + Resnet-34	0.81527
		MLP + EfficientNet-B4	0.88498
IT-ETC	Selected Tabular attributes	Extra Trees Classifier	0.86090
	Selected Tabular attributes and EfficientNet-B4 features		0.88637
IT-PS	**Tabular features and EfficientNet-B4 features**	**MLP + GRU**	**0.90124**

3.5 Image Tabular Extra Tree Classifier Model (IT-ETC)

According to the results from previous related works [1, 21, 30, 35, 40], machine learning models are efficient at most classification problems that use statistical signal-based features as input, especially in human activity classification problems [1, 21, 35]. Based on these works, we deployed different machine learning models such as logistic regression multi-class classifier, random forests, decision tree and extra trees classifier (ETC) from the scikit-learn library [2], with related feature selection techniques to fine-tune each model. Each model was trained with a balanced number of samples for each class in both training and validation sets. Due to the small-scaled dataset, K-Fold cross-validation training ($k = 7$) was applied.

Using the tabular data from CSV file, excluding attributes indicating Resnet-101 features with mouse-logging missing values imputed to -1 and corresponding indicators, the ETC model, which contained 250 randomized decision trees, outperformed other models after feature selection process. Since the ETC was overfitted using raw deep image features in section 3.2, the same feature selection process was applied to choose discriminative components in these features before it was concatenated to corresponding selected tabular features for training.

The feature selection step used in our method is an iterative process which computes impurity-based feature importance for each component in the input vector feature, then choosing the cut-off threshold which was set to 0.001 after empirical trial results. This process was repeated until the model stopped overfitting. The overview pipeline of the IT-ETC approach is illustrated in Figure 1.

3.6 Image Tabular Pair-wise Similarity Model (IT-PS)

The model was designed to cope with the scarcity of available labelled data by following a Siamese structure. The model would take the data from two activities as an input and predict if they come from the same activities or not. By doing this pair-wise approach, we managed to create an enormous number of activity pairs from the small data set. The framework of IT-PS was shown in Figure 2.

We designed an Embedder module to learn the mapping of both image and sensory features into one space. A Discriminator was trained to distinguish two different instances of activities based on their respective embedding vectors. For our particular experiment, let $a = (a_1, ..., a_N)$ and b represented the image and sensory features respectively of an activity instance $\mathbf{x}$. These features were given by the BI and BT model as the fixed input for IT-PS model meaning that this advanced method did not include the layers from these 2 baseline models. We designed the IT-PS model in the following way:

Embedder

$$\begin{aligned} i_t &= MLP([a_t, b]) \\ o_1, ..., o_N &= RNN(i_1, ..., i_N) \\ \mathbf{z} &= MLP(o_N) \end{aligned} \tag{1}$$

where $[\cdot, \cdot]$ represented the concatenation of input vectors, MLP and g were nonlinear, multi-layered functions, RNN was a recurrent neural network and $\mathbf{z}$ was the final embedded vector.

Discriminator

$$\begin{aligned} c(\mathbf{x_1}, \mathbf{x_2}) &= MLP(\mathbf{z}(\mathbf{x_1}), \mathbf{z}(\mathbf{x_2})) \\ P(l_{x_1} = l_{x_2}) &= \sigma(c(\mathbf{x_1}, \mathbf{x_2})) \end{aligned} \tag{2}$$

where σ was the sigmoid function, l_{x_1} and l_{x_2} were the predicted activities of $\mathbf{x_1}$ and $\mathbf{x_2}$ accordingly.

Inference To classify an instance of activity, we ran the trained IT-PS model to compare the instance to each sample of the train dataset and calculate the average score of each class. The scores were then ranked and the highest-scoring class would be taken as the final classification.

3.7 Post-processing

As outlined in the task description, the participating teams were required to submit a ranked list of 140 ids (7 subjects x 20 activities) for each activity (20 in total). This list was evaluated by using the mean Average Precision (mAP) score, a measure which increases when the correct predictions have a high ranking. Considering a specific activity, the most intuitive way to work around this was to sort 140 ids in the descending order. However, since each activity was performed once by each subject (in the test set), this meant that a correctly ranked list should contain 7 activities of 7 distinct subjects at the top. Hence, for each activity, we divided 140 ids into 7 blocks, each of which corresponds to one subject. The id with highest prediction score of each block was then put in the top 7, while the rest is presented in the descending order of prediction score. In effect, this procedure leveraged prior knowledge about the test set collection.

[2] https://scikit-learn.org

4 EXPERIMENTS AND DISCUSSION

4.1 Implementation Details

We configured the BI model by adding a fully connected layer with 512 units after the global pooling layer of the backbone (Resnet-34 or EfficientNet-B4). In the BT model, the size of three hidden layers were 1000, 500 and 128 accordingly. Similarly, the number of nodes of that in IT model was chosen to be 128 also. In the IT-PS model, the MLP structure in the Embedder module was a simple single neural network while that in the Discriminator component was a network with three layers. It is worth pointing out that batch normalisation and drop-out regularisation techniques were utilized after each of the fully connected layers with a drop-out rate of 0.5. We decided to use the GRU module with an embedding size of 512 in the RNN part of the Embedder block due to its compact capacity. We chose Adam [23] to be the optimizer for the BI, IT, IT-PS and BT models. The learning rate was set at 0.001 for the first three approaches while that of the latter was 0.005 (respectively). The value of this hyper-parameter in the BT model was higher since it was trained from scratch. This was different with other designs in which the BI model was already trained from ImageNet dataset [6] while the IT and IT-PS models were initialized from the BI model. We also split the labeled data into training and validate proportionally with the ground truth activities.

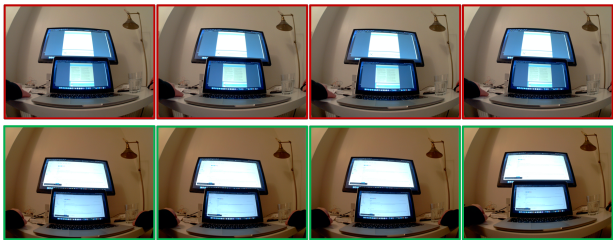

Figure 3: The example of 2 visually similar activities, extracted from subject id 1004. Images highlighted in red were from act02 which is reading text on screen while the ones in green were from act04 which is staring a point

4.2 Results and Discussion

Table 1 provides insight into the mAP results of our submission under different settings. As can be seen from the table, the BT model had the lowest mAP score of 0.698 while IT-PS model performed best with the score more than 0.901. It was not surprising that the baseline structure achieved a lower score than the advanced ones with a difference of at least 0.078 in mAP. In the baseline models, the performance when using images was better than when using non-visual sensor and computer interaction signals. Although containing less number of training parameters, BI model with the backbone of EfficientNet-B4 obtained higher mAP than Resnet-34 which were 0.737 and 0.703 respectively. The difference between the two models became significantly clearer when the score of the IT model using the former backbone was 0.885 while that of the latter was 0.815. This also meant that combining both visual and signal features could boost the performance by up to 0.148 compared to using a single data type (visual vs non-visual). This was anticipated as there were some samples which were impossible to have the correct prediction if only using visual data. For instances, Figure 3 depicted two activities which one would be unable to distinguish by using autographer data alone. The upper row marked with red color illustrates images of the "*Reading text on screen*" activity and the bottom green-noted row illustrates "*Staring*" activity.

The IT-ETC approach employs an ensemble machine learning model that managed to achieve the mAP of 0.861 by using only tabular features processed through a feature selection step. In addition, the approach of concatenating selected EfficientNet-B4 features extracted from autographer increased the mAP slightly to 0.886. In addition, this model attained a higher mAP than IT model on the test set since the low amount of samples in the training dataset is not suitable for a deep learning technique. On the other hand, the IT-PS was the best approach in all runs because of its two main properties. The first was its ability to capture the sequence information between Autographer images within an activity by using a GRU module. The second explanation for its high performance is based on its input format. The IT-PS model received a pair of activity features samples to make the prediction of a probability if they were come from the same activity or not. This meant that by creating pairs of samples it could produce a large number of traiing samples from the training data, hence mitigating issues related to the small training set provided in MART.

However, issues relating to overfitting still remain in our models. Although we split the labeled data into training and validating subsets, the score in the test set is much lower than the mAP in our subset which was almost 1 for the IT model. This matter appears in our baseline BI and BT models, hence also affects the advanced methods which were built on top of two baseline structures. The reason may be the way we created the subset where we randomly selected on entire data. This led to a phenomenon that our training subset included the information of all 7 subjects making it lose some properties of generalisation. One future solution is that we could split the data based on subject-level, for instance, the information of 5 subjects will be used for training and the 2 remaining for validating.

5 CONCLUSION

In this paper, we have presented our team's effort in predicting the micro-activities on the first generation of the MART dataset using both Machine Learning and Deep Learning approaches. A total of five models were proposed to exploit the visual information from the Autographer, signal readings and computer interaction logs to perform the MART retrieval task. The results of two models BI and BT served as baselines for other participating researchers to compare with, while the rest is meant for competitive benchmarking. Among the five models which were used to generate submission runs, we achieved the highest mAP (of 0.90) using our IT-PS model, and overall came second in the MART challenge using this.

For the future work, we aimed to tackle the overfitting issues identified in subsection 4.2 by conducting the user-independent experiments on the proposed models. Additionally, we also plan to engineer more advanced features in future work from the raw data, especially eye movements (i.e. blink rate, fixations, saccades)

and integrate these in our models. Finally, we seek to develop an interactive retrieval system to meaningfully visualise complex multi-modal micro-activity data, that will serve as a search engine as the MART dataset becomes bigger in the future.

6 ACKNOWLEDGMENTS

This publication is funded as part of Dublin City University's Research Committee and research grants from Science Foundation Ireland under grant numbers SFI/12/RC/2289_2, SFI/13/RC/2106, 18/CRT/6223, and 18/CRT/6224.

REFERENCES

[1] Muhammad Sufyian Mohd Azmi and Nasir Sulaiman. 2017. Accelerator-Based Human Activity Recognition Using Voting Technique with NBTree and MLP Classifiers. *International Journal on Advanced Science, Engineering and Information Technology* 7 (2017), 146–152.

[2] A. Bulling, J. A. Ward, H. Gellersen, and G. Tröster. 2011. Eye Movement Analysis for Activity Recognition Using Electrooculography. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 33, 4 (2011), 741–753.

[3] Duc-Tien Dang-Nguyen, Luca Piras, Michael Riegler, Giulia Boato, Liting Zhou, and Cathal Gurrin. 2017. Overview of ImageCLEFlifelog 2017: Lifelog Retrieval and Summarization. In *CLEF*.

[4] Duc-Tien Dang-Nguyen, Luca Piras, Michael Riegler, Liting Zhou, Matthias Lux, and Cathal Gurrin. 2018. Overview of ImageCLEFlifelog 2018: Daily Living Understanding and Lifelog Moment Retrieval. In *CLEF*.

[5] Duc-Tien Dang-Nguyen, Luca Piras, Michael Riegler, Liting Zhou, Matthias Lux, Minh-Triet Tran, Tu-Khiem Le, Van-Tu Ninh, and Cathal Gurrin. 2019. Overview of ImageCLEFlifelog 2019: Solve My Life Puzzle and Lifelog Moment Retrieval. In *CLEF*.

[6] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 248–255.

[7] Aaron Duane, Björn Þór Jónsson, and Cathal Gurrin. 2020. VRLE: Lifelog Interaction Prototype in Virtual Reality: Lifelog Search Challenge at ACM ICMR 2020. In *Proceedings of the Third Annual Workshop on Lifelog Search Challenge (LSC '20)*. Association for Computing Machinery, New York, NY, USA, 7–12.

[8] Cathal Gurrin, Hideo Joho, Frank Hopfgartner, Liting Zhou, and Rami Albatal. 2016. Overview of NTCIR-12 Lifelog Task. In *NTCIR*.

[9] Cathal Gurrin, Hideo Joho, Frank Hopfgartner, Liting Zhou, Rashmi Gupta, Rami Albatal, and Duc-Tien Dang-Nguyen. 2017. Overview of NTCIR-13 Lifelog-2 Task.

[10] Cathal Gurrin, Hideo Joho, Frank Hopfgartner, Liting Zhou, Van-Tu Ninh, Tu-Khiem Le, Rami Albatal, Duc-Tien Dang-Nguyen, and Graham Healy. 2019. Overview of the NTCIR-14 Lifelog-3 task.

[11] Cathal Gurrin, Tu-Khiem Le, Van-Tu Ninh, Duc-Tien Dang-Nguyen, Björn Þór Jónsson, Jakub Lokoš, Wolfgang Hürst, Minh-Triet Tran, and Klaus Schöffmann. 2020. Introduction to the Third Annual Lifelog Search Challenge (LSC'20). In *Proceedings of the 2020 International Conference on Multimedia Retrieval (ICMR '20)*. Association for Computing Machinery, New York, NY, USA, 584–585.

[12] Cathal Gurrin, Klaus Schoe, Mann, Hideo Joho, Andreas Leibetseder, Liting Zhou, Aaron Duane, Duc-Tien Dang-Nguyen, Michael Alexander Riegler, Luca Piras, Minh-Triet Tran, Jakub Lokoš, and Wolfgang Hürst. 2019. Paper Comparing Approaches to Interactive Lifelog Search at the Lifelog Search Challenge (LSC 2018).

[13] Cathal Gurrin, Klaus Schoeffmann, Hideo Joho, Andreas Leibetseder, Liting Zhou, Aaron Duane, Duc-Tien Dang-Nguyen, Michael Riegler, Luca Piras, Minh-Triet Tran, Jakub Lokoč, and Wolfgang Hürst. 2019. [Invited papers] Comparing Approaches to Interactive Lifelog Search at the Lifelog Search Challenge (LSC2018). *ITE Transactions on Media Technology and Applications* 7, 2 (2019), 46–59.

[14] Cathal Gurrin, Alan F. Smeaton, and Aiden R. Doherty. 2014. LifeLogging: Personal Big Data. *Found. Trends Inf. Retr.* 8, 1 (June 2014), 1–125.

[15] M. Hassan, M. Z. Uddin, A. Mohamed, and Ahmad Almogren. 2018. A robust human activity recognition system using smartphone sensors and deep learning. *Future Gener. Comput. Syst.* 81 (2018), 307–313.

[16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015. Deep Residual Learning for Image Recognition. arXiv:cs.CV/1512.03385

[17] Graham Healy, Tu-Khiem Le, Hideo Joho, Frank Hopfgartner, and Cathal Gurrin. 2020. Overview of NTCIR-15 MART. In *Proceedings of the NTCIR-15 Conference*. Tokyo, Japan.

[18] Silvan Heller, Mahnaz Amiri Parian, Ralph Gasser, Loris Sauter, and Heiko Schuldt. 2020. Interactive Lifelog Retrieval with Vitrivr. In *Proceedings of the Third Annual Workshop on Lifelog Search Challenge (LSC '20)*. Association for Computing Machinery, New York, NY, USA, 1–6.

[19] Andrey Ignatov. 2018. Real-time human activity recognition from accelerometer data using Convolutional Neural Networks. *Appl. Soft Comput.* 62 (2018), 915–922.

[20] Shoya Ishimaru, Kai Kunze, Yuji Uema, Koichi Kise, Masahiko Inami, and Katsuma Tanaka. 2014. Smarter Eyewear: Using Commercial EOG Glasses for Activity Recognition. In *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct Publication (UbiComp '14 Adjunct)*. Association for Computing Machinery, New York, NY, USA, 239–242.

[21] Ankita Jain and Vivek Kanhangad. 2018. Human Activity Classification in Smartphones Using Accelerometer and Gyroscope Sensors. *IEEE Sensors Journal* 18 (2018), 1169–1177.

[22] Adil Khan, Muhammad Siddiqi, and Seok-Won Lee. 2013. Exploratory Data Analysis of Acceleration Signals to Select Light-Weight and Accurate Features for Real-Time Activity Recognition on Smartphones. *Sensors (Basel, Switzerland)* 13 (2013), 13099 – 13122.

[23] Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).

[24] Mustafa Kose, Ozlem Incel, and Cem Ersoy. 2012. Online Human Activity Recognition on Smart Phones. *Workshop on Mobile Sensing: From Smartphones and Wearables to Big Data* (01 2012).

[25] Gregor Kovalčík, Vít Škrhak, Tomáš Souček, and Jakub Lokoč. 2020. VIRET Tool with Advanced Visual Browsing and Feedback. In *Proceedings of the Third Annual Workshop on Lifelog Search Challenge (LSC '20)*. Association for Computing Machinery, New York, NY, USA, 63–66.

[26] Yunji Liang, Xingshe Zhou, Z. Yu, and Bin Guo. 2014. Energy-Efficient Motion Related Activity Recognition on Mobile Devices for Pervasive Healthcare. *Mobile Networks and Applications* 19 (2014), 303–317.

[27] František Mejzlík, Patrik Veselý, Miroslav Kratochvíl, Tomáš Souček, and Jakub Lokoč. 2020. SOMHunter for Lifelog Search. In *Proceedings of the Third Annual Workshop on Lifelog Search Challenge (LSC '20)*. Association for Computing Machinery, New York, NY, USA, 73–75.

[28] Van-Tu Ninh, Tu-Khiem Le, Liting Zhou, Luca Piras, Michael Riegler, Pål Halvorsen, Minh-Triet Tran, Mathias Lux, Cathal Gurrin, and Duc-Tien Dang-Nguyen. 2020. Overview of ImageCLEF Lifelog 2020:Lifelog Moment Retrieval and Sport Performance Lifelog. In *CLEF2020 Working Notes (CEUR Workshop Proceedings)*. CEUR-WS.org <http://ceur-ws.org>, Thessaloniki, Greece.

[29] Francisco Ordóñez and Daniel Roggen. 2016. Deep Convolutional and LSTM Recurrent Neural Networks for Multimodal Wearable Activity Recognition. *Sensors* 16, 1 (Jan 2016), 115.

[30] Saroj Kumar Panigrahy, Sanjay Kumar Jena, , and Ashok Kumar Turuk. 2017. Study and Analysis of Human Stress Detection using Galvanic Skin Response (GSR) Sensor in Wired and Wireless Environments. *Research Journal of Pharmacy and Technology* 10 (2017), 545–550.

[31] Muhammad Shoaib, S. Bosch, Özlem Durmaz Incel, H. Scholten, and P. Havinga. 2014. Fusion of Smartphone Motion Sensors for Physical Activity Recognition. *Sensors (Basel, Switzerland)* 14 (2014), 10146 – 10176.

[32] Pekka Siirtola, Heli Koskimäki, and J. Röning. 2016. From User-independent to Personal Human Activity Recognition Models Using Smartphone Sensors. In *ESANN*.

[33] Pekka Siirtola and J. Röning. 2013. Ready-to-use activity recognition for smartphones. *2013 IEEE Symposium on Computational Intelligence and Data Mining (CIDM)* (2013), 59–64.

[34] Monit Shah Singh, Vinaychandran Pondenkandath, Bo Zhou, P. Lukowicz, and Marcus Liwicki. 2017. Transforming sensor data to the image domain for deep learning — An application to footstep detection. *2017 International Joint Conference on Neural Networks (IJCNN)* (2017), 2665–2672.

[35] Abdul Syafiq Abdull Sukor, Ammar Zakaria, and N. H. A. Rahim. 2018. Activity recognition using accelerometer sensor and machine learning classifiers. *2018 IEEE 14th International Colloquium on Signal Processing & Its Applications (CSPA)* (2018), 233–238.

[36] M. Tan and Quoc V. Le. 2019. EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. *ArXiv* abs/1905.11946 (2019).

[37] Ly-Duyen Tran, Manh-Duy Nguyen, Nguyen Thanh Binh, Hyowon Lee, and Cathal Gurrin. 2020. Myscéal: An Experimental Interactive Lifelog Retrieval System for LSC'20. In *Proceedings of the Third Annual Workshop on Lifelog Search Challenge*. 23–28.

[38] Jindong Wang, Y. Chen, Shuji Hao, Xiaohui Peng, and L. Hu. 2019. Deep Learning for Sensor-based Activity Recognition: A Survey. *ArXiv* abs/1707.03502 (2019).

[39] Shuochao Yao, Shaohan Hu, Yiran Zhao, Aston Zhang, and Tarek Abdelzaher. 2017. DeepSense: A Unified Deep Learning Framework for Time-Series Mobile Sensing Data Processing. In *Proceedings of the 26th International Conference on World Wide Web (WWW '17)*. International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, CHE, 351–360.

[40] P. Zontone, A. Affanni, R. Bernardini, A. Piras, and R. Rinaldo. 2019. Stress Detection Through Electrodermal Activity (EDA) and Electrocardiogram (ECG) Analysis in Car Drivers. *2019 27th European Signal Processing Conference (EUSIPCO)* (2019), 1–5.