Improved Lifelog Ego-centric Video Summarization Using Ensemble of Deep Learned Object Features

Philip Mainwaring
philipjmainwaring@gmail.com
Bappaditya Mandal
https://www.keele.ac.uk/scm/staff/drbappadityamandal/

School of Computing and Mathematics
Keele University
Staffordshire ST5 5BG, UK

Abstract

The ImageCLEF 2017 lifelog summarization challenge [10, 12] was established to develop a benchmark for summarizing egocentric lifelogging videos based on our daily activities, such as 'commute to work' or 'cooking at home'. In this paper, we propose an iterative approach for summarizing lifelogging activities based on task queries provided by the ImageCLEF 2017 lifelog summarization challenge. YoloV3 image detection, TensorFlow GoogleNet image classification and Places365 environment classification resources are used to generate low level deep learned features from the lifelogging images. A nearest neighbor classifier is used to generate high level descriptors to classify lifelogger activities per image basis, which is also a requirement as provided in the ground truth labels. Finally, key frame images per activity are selected via hierarchical clustering to create an accurate and diverse static storyboard of summarized lifelog activities. Experimental results show the superiority of the proposed approach as compared to the highest reported results achieved in the ImageCLEF 2017 lifelog summarization competition.

1 Introduction

The availability of devices such as the Narrative Clip or GoPro cameras [9] and many other wearable devices, allow visual egocentric recording of a user's everyday life. The Narrative Clip for instance, can be attached to a user's chest (egocentric view) and record one picture per minute of their daily life. This personal media archive contains vast amounts of data collected from minute by minute recordings captured over the course of months. This visual log (lifelog) is outlined by Sellen and Whittaker [26] as a key component in maintaining a personal archive of augmented memory. Augmenting a user's memory has many benefits such as recollecting events, reflecting / reminiscing on past experiences and retrieving information such as the last location of a lost object or recognizing an individual [3, 18] or retrieving episodic memories [4] or recognizing ego-centric activities [27, 28]. There is an increasing demand for techniques to summarize these archives of personal big data, allowing data to be efficiently stored, analyzed and retrieved. Molino *et al.* in [5] presented a recent survey on summarizing the ego-centric videos. It also reports recreational and occupational applications of lifelogging, such as recording special life events and extreme experiences.

Police officers can record a patrol route, these recordings are then used as evidence. Selke [25] lists the capacity for surveillance and detecting dangerous situations for soldiers in hostile territory. Monitoring caregivers' supervision and medicine administration. Tracking patient routines can lead to better diagnosis and personalized care giving. Xu *et al.* [31] developed a wearable system to remind and monitor users taking medicine. This is specifically advantageous for dementia patients who may not remember taking their medicine. Recording and describing instructional advice videos and providing user-friendly walkthroughs are a subset of social media. Kerr *et al.* [15] developed an arm-mounted augmented reality system to assist navigation in an outdoor environment. Tracking these lifelogging activities and permanently storing images as a personal media archive produces more data than can be manually categorized and summarized without excessive investment of time and effort by the user. This has led to an increasing demand for techniques to summarize these archives of personal big data [8].

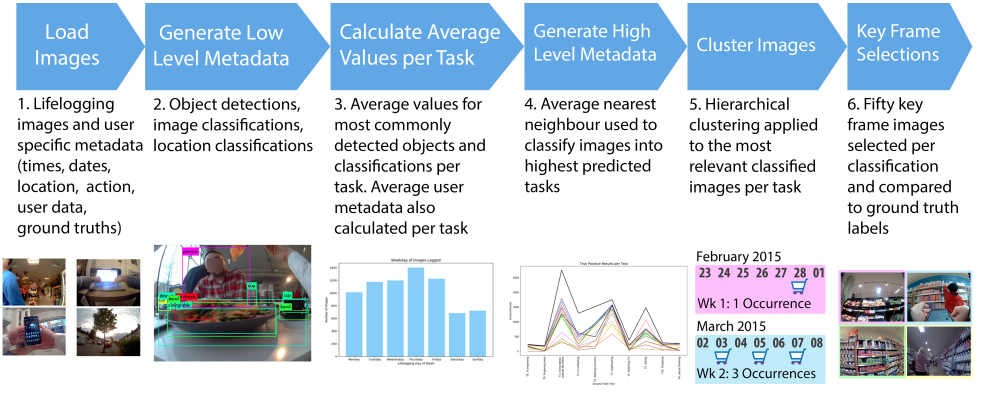

Figure 1: The proposed model architecture where average nearest neighbor classifier along with hierarchical clustering are used for key frame selection (best viewed in colour and zoomed in).

2 Related Work

The annual ImageCLEF competition [12] addressed the lack of retrieval and summarization techniques by launching a competition to develop a new benchmark in lifelog categorization and evaluation. Details of the task (SubTask 2: Lifelog summarization (LST)) can be found in the website [10].

2.1 ImageCLEF 2017 Challenge Dataset, Problem Statement, Requirements and Results

The challenge involved summarizing month long lifelogs of three lifeloggers into activity categorizations based on specific requirements. Participants were provided with 88,124 lifelogging images (roughly 1-2 frames per minute) [10]. The challenge was to analyze and correctly identify which images corresponded to predefined activity categorizations. A

development set of images relating to five ground truth categories was provided to allow participants to train their models. Testing was independently performed, allowing the retrieval of a set of key frames from the database that would summarize each of the 10 activities (test queries) to represent 10 diverse and accurate storyboard.

For example a query in the development dataset [10, 11]:
Shopping: Summarize the moment(s) in which user u1 doing shopping.
Description: To be relevant, the user must clearly be inside a supermarket or shopping stores (includes book store, convenient store, pharmacy, etc). Passing by or otherwise seeing a supermarket are not considered relevant if the user does not enter the shop to go shopping. Blurred or out of focus images are not relevant. Images that are covered (mostly by the lifelogger's arm) are not relevant.

Other interesting queries can be found in the database of [11]. Participants in this lifelog video summarization have used average F1-score at $X = 10$ (number of retrieved images) of the ten activity categorizations to be summarized. X would vary between 5 and 50, $X : X \in \{5, 10, 20, 30, 40, 50\}$. The highest achieving results in the competition were accomplished by Dogariu and Ionescu (UPB) [7] with a 0.132 F1-score accuracy and Molino *et al.* (I2R) [19] with a 0.497 F1-score accuracy. Dogariu *et al.* [7] combined visual and textual data into written attributes per image and assessed word similarity to cluster the data and summarize the images. This is not seen as a successful approach, concluding that image object detectors customized to the lifelogging activities and common objects would improve results.

Molino *et al.* [19] combined image metadata and lifelogger-provided data (e.g. locations) to extract the parameters for clustering and training interactive machine learning. Their conclusion is that extra user-provided metadata and task-specific techniques rather than a generic summarization approach would improve results. Overall, the challenge failed to establish a benchmark for lifelog summarization. Sixteen groups participated in this challenge, however only two managed to submit results. An overview of the challenge results [20] suggests the complexity of the task and the difficulty level of data to be processed as the most probable reasons for this failure. This implies that for this well archived database and well defined problem, attention from a larger research community is required.

2.2 Aim

Since the accuracy obtained by the best performing algorithms is low, in this work, we aim to develop a better algorithm that could outperform all the existing reported works, such as F1-score of 0.497 obtained in [19]. We also plan to automate the model, removing the need for task-specific weighting or structured rules to improve categorization accuracy reported in [19]. To achieve this aim, YoloV3 image detection, TensorFlow GoogleNet image classification and Places365 environment classification resources have been used to detect objects and subsequently, perform object and place recognitions. Thereby, creating the low level deep learned features. An average-nearest-neighbor classifier is proposed to establish activity specific average object prediction values, and hierarchical clustering is used to select n appropriate key frames per activity task.

3 Proposed Architecture

The proposed architecture is shown in Fig. 1, each stage is explained below.

3.1 Automatic Extraction of Low Level Deep Learned Features

Appropriate object identification is reported as a key factor in the highest achieving models in the competition [19]. Being able to accurately identify many objects provided more evidence to justify an activity classification than if only a few objects are identified. Image object detectors and classifiers are used to enrich the descriptive features. YoloV3 image detection [22, 23] is used with the MS Coco dataset [17]; this resource is well established for object detection and has 80 object classes. YoloV3 image detection does not have highest level of accuracy of the models tested, however, the speed of predictions with good accuracy allowed detections to be completed within a reasonable time frame. It took roughly 11 days to process 88,124 lifelogging images with the YoloV3 model on a machine with an Intel core i7-3770 CPU, 3.40GHz with 8GB of RAM. If Faster R-CNN (region-convolutional neural networks) model [24] was used, this could have been much longer.

For image object classification, one of the highest performing deep convolutional neural networks (CNN) is the TensorFlow Inception-v3 network [29, 30], which has a very low prediction error rate. Although this network is not the latest version, it is reliably well established and frequently used in the ImageNet challenges. Tensorflow GoogleNet image classifier [1] is used with the ImageNet Dataset [6]. The ImageNet dataset contains 21,841 object classes and sets the benchmark for image classification. Some sample detections and classifications are shown in Supp A in the supplementary accompanied with this paper. Image classifiers have also been modified to recognition tasks. For example, the scene recognition 'Places365' [33] attempts to classify the image environment (e.g. shoe shop, cafeteria, bedroom) based on a dataset of 434 classes. Likewise, action recognition resources [16, 32] attempt to classify the predominant action occurring in an image (e.g. gymnastics, cricket). Collating image attributes from a collection of these resources provides a rich source of descriptive features allowing the lifelogging images to be classified.

Images from wearable devices and user-specific metadata (e.g. image time and date) provided by ImageCLEF in an XML document are stored in a Python dictionary. The values from each identifier are collated as image attributes in the central Python dictionary. Bounding boxes used to show the coordinates of object detections in the YoloV3 results are used to identify whether an object is in the foreground (large) or in the background (small) as shown in Fig. 1. This is added to assist categorizations such as social drinking where the lifelogger needed to be drinking with people rather than just having other people present in the room.

3.2 Calculating Average Values per Task

Since the images are captured using wearable devices by human participants in our common daily life routine, both the target and capturing devices were moving continuously, resulting in many poor quality images, such as blurry and out-of-focus images. The image ground truths (10,137 correctly labeled images) were made available following the ImageCLEF 2017 lifelog summarization challenge. This allows a classifier to learn effectively from the labeled images using the development dataset. The first step is to identify the most common image predictions for each activity. A 'perfect average' of attribute likelihoods is established to be compared against. Many careful analysis of the images in the development dataset using Laplacian variance, average blurriness by image segment, temporal aspects (course of lifelog) of the image captured, including the days and times in a week and the location of capturing are performed systematically for each of the tasks (development queries) to get better understanding of the egocentric lifelog videos. Extended analysis for detections

and classifications are performed for Tensorflow ImageNet classifications for 10 most common objects. Similar experiments are repeated for 10 most common objects using YoloV3 (objects) and Places365 environment (location) classifications. Finally experiments are also conducted for finding ten most common DarkNet ImageNet classifications for each of development tasks. Detail analysis of the average values per task (queries) for developmental and test query sets are provided in Supp B in the supplement of this paper.

3.3 Automatic Extraction of High Level Descriptors

A python based average-nearest-neighbor classifier compares each new image to the 'perfect average' identification attributes per task. An average knowledge-based classifier is preferred to a *k*-nearest neighbor classifier due to the nature of images being labeled. For instance if a lifelogger labeled an hour of images as 'In a Meeting', and for a small part of this hour the lifelogger needed to use the toilet, then any subsequent visit to a toilet could be labeled as being 'In a Meeting'. Using an average of image attributes for this hour limits this possibility. The classification likelihood is attributed to each image as part of the feature descriptors. Leave-one-out testing is applied to ground truth labeled images, average results are shown in Table 1. These images individually have their ground truth labels removed before the classifier predicts the most likely activity.

This also allows for an increased personalization of the lifelogging categorizations. The most popular images to occur during a meeting are planetariums and woks (due to a circular ceiling light being incorrectly identified). While this is obviously inaccurate, it is also replicable. If a new image is added and the two most prominent objects classified are a planetarium and a wok, then the lifelogger is most likely in a meeting (also illustrated by example images in the Supp C in the supplement of this paper). This demonstrates a key difference between this study and the main entries in the Lifelog competition. By attempting to define human understandable logical rules to image object recognition, the main elements identified in the image will not register. Whereas if the model is allowed to generate its own rules for what is required for a classification, then a higher level of personalized predictions can be achieved. This leads to a higher level of personalization and more accurate predictions for each user.

3.4 Clustering Images

The most confidently predicted images for each classification are selected to ensure the key frames are relevant. Hierarchical clustering is then used to select key frames representing *relevance* and *diversity* of each activity. In the ImageCLEF 2017 lifelog summarization challenge, *relevance* is similarity among the retrieved images with respect to the given task or query. *Diversity* implied the retrieved image set should be comprising of images from various times and days, considering the dissimilarity between the individual items in the general content of the images. Apart from detected and recognized objects, places or landmarks and other attributes, this incorporates an aesthetic evaluation [2] allowing clusters to include aspects such as average hue and color distribution.

3.5 Key Frame Selection

Selected key frame images are then listed with predicted and actual ground truth labels (if present) to test the accuracy of the model. The precision, recall and F1-scores are calculated

by investigating the number of correctly and incorrectly predicted activities. The prediction accuracy improved as the number of clustered images available for key frame selection are reduced. A manual inspection showed the image diversity also reduced along with the number of images. Experimentation revealed 'three times the number of key frame selections' provided the best number of clustered images for key frame selection relevance and diversity.

Table 1: Average F1-Score Results for $X = 10$, where X is the number of retrieved images from the dataset.

Methods	Images Only	Images & User-Entered Metadata
ImageCLEF 2017 highest ranking results	0.132 UPB [7]	0.497 I2R [19]
Leave-one-out testing	0.749	0.782
Our Proposed Method	**0.688**	**0.631**

One aspects given less attention in the ImageCLEF competition (2017a) was aesthetic values and quality of the image. The image quality was added to the image descriptors allowing for a better clustering algorithm to be built. Use of a dendrogram to evaluate the clustering accuracy, permitted a weighting of key frame attributes and classifications to show which values provided the most accurate results. Greater weighting were given to the Places365 and the TensorFlow classifications. This also allowed for re-appraisal and improved examination of classifications such as the Yolo results; bounding boxes were analyzed to compare whether objects (especially people) were closer to the lifelogger. Weighting was also applied to the image descriptors in run 2 of the experiments.

4 Experimental Results and Discussions

The performance evaluation metrics are defined by the organizers [10, 17]:

1. Cluster Recall at X ($CR@X$), where X is the number of retrieved images from the database: a metric that assesses how many different clusters from the ground truth are represented among the top X results (*diversity* measure);

2. Precision at X ($P@X$): measures the number of relevant photos among the top X results (measure of *relevance*);

3. F1-measure at X ($F1@X$): the harmonic mean of the previous two. Takes both into account the *diversity* and *relevance*.

Official ranking metrics in the year 2017 happened to be the $F1-measure@X = 10$, which gives equal importance to diversity (via $CR@X = 10$) and relevance (via $P@X = 10$). Further details are provided in [11].

4.1 Results

The results are divided into two runs (image analysis only and image analysis with user-entered metadata). Table 1 shows our results compared to the highest ranking ImageCLEF results as well as the average F1-Scores from the leave-one-out testing explained earlier. The

results exceed the ImageCLEF challenge results, achieving the goal of surpassing average F1-scores for each run. The results improve when user-provided metadata is not used, in activities other than when lifelogger 1 is in a meeting, demonstrating the model is not dependent on the lifelogger manually adding metadata. This is less noticeable as the number of key frames selected increases (Fig. 2).

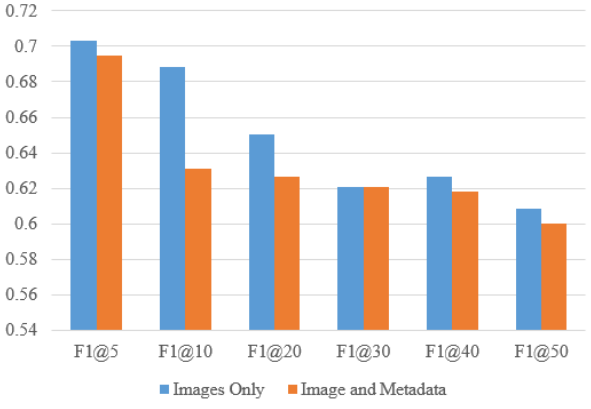

Figure 2: Average F1-score at X={5, 10, 20, 30, 40, 50} (such as F1@10), as key frame selections increase: x-axis shows the number of key frames selected, y-axis shows the F1-score values (in the range of 0 to 1). The results start to coincide as the number of key frames increase (best viewed in colour).

Fig. 2 shows the average F1-score at various number of top retrieved images X from the database considering both the 'relevance' and 'diversity' [10], where $X = 5, 10, 20, 30, 40, 50$. A downward curve occurs in Fig. 2, because the model selects n of the most confident predictions available. The model includes images with lower confidence scores, producing lower overall accuracy, as more selections are made. Further investigation also revealed the day and time data can provide misleading predictions. If a lifelogger passes a bus at an usual time when they commute, this may be incorrectly categorized with a high prediction confidence.

Fig. 3 shows the F1-score at various X for all the 10 test tasks (or test queries [10]). It can be observed that for some tasks, such as 'Transporting' or 'shopping' the performance could be high, but for other tasks, such as 'working at home' or 'in a meeting', the accuracies are low. This probably shows that depending on the difficulty level of the tasks, performances could vary significantly. Our incremental development focused on improving the accuracy of the model while reducing the demand for the user to enter data (images only). The only input required from the lifelogger is to accurately label a development sample of images. There is no need for complex rules for structured learning as evident in the model by Molino *et al.* [19], which requires human understandable logical rules for activity categorization. Through an automated approach, personalization and accuracy of the model improves as the development set of ground truth labels increases. More experimental results and analysis are presented in the Supp D of the supplementary document. As shown in Table 1, our proposed approach outperforms all the previously reported results on this egocentric video summarization task.

Regarding key frame diversity, improvements could be made to the range of images se-

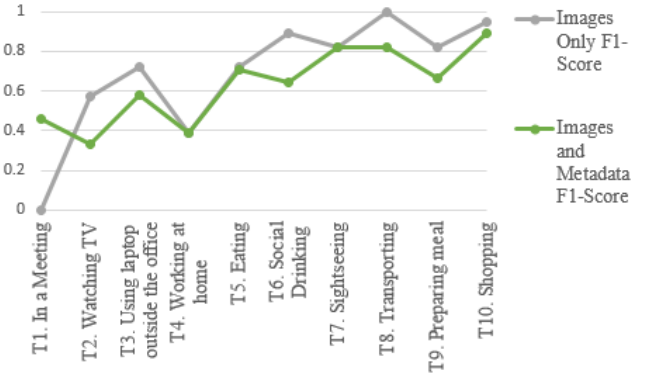

Figure 3: Average F1-score at X=10 (F1@10), accuracy results: the x-axis shows the activity categorizations, the y-axis shows F1-score. Image only values are marginally better than the image and metadata values (best viewed in color).

lected. This could occur from better use of the user supplied metadata such as time and location. The primary intent in the experimentation was to deliver accurate results rather than diverse results as these are easier to measure. On some occasions, images selected were within a few minutes of similar images. This would need to be amended for the summarization to accurately portray category diversity.

5 Conclusions and Future Work

In this work, we have proposed a framework for processing egocentric lifelog videos captured by 3 lifeloggers for a month long resulting in 88,124 images provided by the organizers of ImageCLEF 2017 lifelog summarization challenge. In our proposed approach, the image metadata did not have a significant impact on the results showing that the model is capable of making accurate predictions without input from the user other than development ground truth labels. Our framework does not require correct image object predictions to classify images also it avoids task-specific weighting or structured rules to improve the categorization accuracy (as done in the current state-of-the-art [19]). Hence, our approach could accommodate improvements in image identification resources, such as Yolo9000 [21] or Kaggle ImageNet Object Localization Challenge [14], as these provide further data for classifications. An average-nearest-neighbor classifier proved to be a useful addition as it allowed analysis of the most common values to be identified and it avoided results being adversely affected by outlier variables. Our proposed approach achieved the accuracy which outperformed the highest rating submissions for the SubTask 2: Lifelog summarization in ImageCLEF 2017 lifelog summarization challenge [10]. This approach of classifying and clustering images for static storyboard key frame selection could also be used for the latest ImageCLEF lifelog challenges [13]. If it is used to generate contextual data per image, the model can also be used to improve other summarization techniques such as dynamic video skimming.

References

[1] Martín Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard, et al. Tensorflow: a system for large-scale machine learning. In *OSDI*, volume 16, pages 265–283, 2016.

[2] E. Charles. photo-quality. https://github.com/dresa/photo-quality, 2013.

[3] Shue-Ching Chia, Bappaditya Mandal, Qianli Xu, Liyuan Li, and Joo-Hwee Lim. Enhancing social interaction with seamless face recognition on google glass: Leveraging opportunistic multi-tasking on smart phones. In *Proceedings of the 17th International Conference on Human-Computer Interaction with Mobile Devices and Services Adjunct, MobileHCI '15, Copenhagen, Denmark, August 24-27, 2015*, pages 750–757, 2015.

[4] Ana Garcia del Molino, Bappaditya Mandal, Liyuan Li, and Joo-Hwee Lim. Organizing and retrieving episodic memories from first person view. In *2015 IEEE International Conference on Multimedia & Expo Workshops, ICME Workshops 2015, Turin, Italy, June 29 - July 3, 2015*, pages 1–6, 2015.

[5] Ana Garcia del Molino, Cheston Tan, Joo-Hwee Lim, and Ah-Hwee Tan. Summarization of egocentric videos: A comprehensive survey. *IEEE Transactions on Human-Machine Systems*, 47(1):65–76, 2017.

[6] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 248–255. Ieee, 2009.

[7] Mihai Dogariu and Bogdan Ionescu. A textual filtering of hog-based hierarchical clustering of lifelog data. *CLEF working notes, CEUR (September 11-14 2017)*, 2017.

[8] Aaron Duane, Rashmi Gupta, Liting Zhou, and Cathal Gurrin. Visual insights from personal lifelogs. In *Proceedings of the 12th NTCIR Conference on Evaluation of Information Access Technologies, Tokyo*, pages 386–389, 2016.

[9] GoPro. Wearable device. http://gopro.com/, 2018.

[10] ImageCLEFlifelog. Subtask 2: Lifelog summarization (lst). https://www.imageclef.org/2017/lifelog, 2019.

[11] ImageCLEFlifelogDataset. Imageclef 2017 - lifelog task - getting datasets. http://imageclef-lifelog.computing.dcu.ie/2017/, 2019.

[12] Bogdan Ionescu, Henning Müller, Mauricio Villegas, Helbert Arenas, Giulia Boato, Duc-Tien Dang-Nguyen, Yashin Dicente Cid, Carsten Eickhoff, Alba G Seco de Herrera, Cathal Gurrin, et al. Overview of imageclef 2017: Information extraction from images. In *International Conference of the Cross-Language Evaluation Forum for European Languages*, pages 315–337. Springer, 2017.

[13] Bogdan Ionescu, Henning Müller, Mauricio Villegas, Alba García Seco de Herrera, Carsten Eickhoff, Vincent Andrearczyk, Yashin Dicente Cid, Vitali Liauchuk, Vassili Kovalev, Sadid A Hasan, et al. Overview of imageclef 2018: Challenges, datasets and evaluation. In *International Conference of the Cross-Language Evaluation Forum for European Languages*, pages 309–334. Springer, 2018.

[14] Kaggle. Imagenet object localization challenge. https://www.kaggle.com/c/imagenet-object-localization-challenge, 2018.

[15] Steven J Kerr, Mark D Rice, GT Jackson Lum, and Marcus Wan. Evaluation of an arm-mounted augmented reality system in an outdoor environment. In *Network of Ergonomics Societies Conference (SEANES), 2012 Southeast Asian*, pages 1–6. IEEE, 2012.

[16] Tianwei Lin, Xu Zhao, and Zheng Shou. Single shot temporal action detection. In *Proceedings of the 25th ACM International Conference on Multimedia*, MM '17, pages 988–996, 2017. ISBN 978-1-4503-4906-2.

[17] TY Lin, M Maire, S Belongie, J Hays, P Perona, D Ramanan, P Dollár, and CL Zitnick. Microsoft COCO: Common objects in context. ineuropean conference on computer vision 2014 sep 6 (pp. 740-755).

[18] Bappaditya Mandal, Shue-Ching Chia, Liyuan Li, Vijay Chandrasekhar, Cheston Tan, and Joo-Hwee Lim. A wearable face recognition system on google glass for assisting social interactions. In *Computer Vision - ACCV 2014 Workshops - Singapore, Singapore, November 1-2, 2014*, pages 419–433, 2014.

[19] Ana Garcia del Molino, Bappaditya Mandal, Jie Lin, Joo Hwee Lim, Vigneshwaran Subbaraju, and Vijay Chandrasekhar. VC-I2R@ imageclef2017: Ensemble of deep learned features for lifelog video summarization. 2017.

[20] Dang Nguyen, Duc Tien, Luca Piras, Michael Riegler, Giulia Boato, Liting Zhou, and Cathal Gurrin. Overview of imageclef lifelog 2017: lifelog retrieval and summarization. 2017.

[21] Joseph Redmon and Ali Farhadi. Yolo9000: better, faster, stronger. *arXiv preprint*, 2017.

[22] Joseph Redmon and Ali Farhadi. Yolov3: An incremental improvement. *CoRR*, abs/1804.02767, 2018. URL http://arxiv.org/abs/1804.02767.

[23] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 779–788, 2016.

[24] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 1*, NIPS'15, pages 91–99, Cambridge, MA, USA, 2015.

[25] Stefan Selke. *Lifelogging: Digital self-tracking and Lifelogging-between disruptive technology and cultural transformation*. Springer, 2016.

[26] Abigail J Sellen and Steve Whittaker. Beyond total capture: a constructive critique of lifelogging. *Communications of the ACM*, 53(5):70–77, 2010.

[27] Sibo Song, Vijay Chandrasekhar, Bappaditya Mandal, Liyuan Li, Joo-Hwee Lim, Giduthuri Sateesh Babu, Phyo Phyo San, and Ngai-Man Cheung. Multimodal multi-stream deep learning for egocentric activity recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops 2016, Las Vegas, NV, USA, June 26 - July 1, 2016*, pages 378–385, 2016.

[28] Sibo Song, Ngai-Man Cheung, Vijay Chandrasekhar, Bappaditya Mandal, and Jie Lin. Egocentric activity recognition with multimodal fisher vector. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2016, Shanghai, China, March 20-25, 2016*, pages 2717–2721, 2016.

[29] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 2818–2826, 2016.

[30] Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alexander A. Alemi. Inception-v4, inception-resnet and the impact of residual connections on learning. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, February 4-9, 2017, San Francisco, California, USA.*, pages 4278–4284, 2017.

[31] Qianli Xu, Shue Ching Chia, Joo-Hwee Lim, Yiqun Li, Bappaditya Mandal, and Liyuan Li. Medhelp: enhancing medication compliance for demented elderly people with wearable visual intelligence. *Scientific Phone Apps and Mobile Devices*, 2(1): 3, 2016.

[32] Yue Zhao, Yuanjun Xiong, Limin Wang, Zhirong Wu, Xiaoou Tang, and Dahua Lin. Temporal action detection with structured segment networks. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, pages 2933–2942, 2017.

[33] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Antonio Torralba, and Aude Oliva. Places: An image database for deep scene understanding. *arXiv preprint arXiv:1610.02055*, 2016.