DO YOU LIVE A HEALTHY LIFE? ANALYZING LIFESTYLE BY VISUAL LIFE LOGGING

Qing Gao Mingtao Pei Hongyu Shen

Beijing Institute of Technology
Beijing Laboratory of Intelligent Information Technology
Beijing 100081, P.R. China

ABSTRACT

In this work, we investigate the problem of lifestyle analysis and build a visual lifelogging dataset for lifestyle analysis (VLDLA). The VLDLA contains images captured by a wearable camera every 3 seconds from 8:00 am to 6:00 pm for seven days. In contrast to current lifelogging/egocentric datasets, our dataset is suitable for lifestyle analysis as images are taken with short intervals to capture activities of short duration; moreover, images are taken continuously from morning to evening to record all the activities performed by a user. Based on our dataset, we classify the user activities in each frame and use three latent fluents of the user, which change over time and are associated with activities, to measure the healthy degree of the user's lifestyle. Experimental results show that our method can be used to analyze the healthiness of users' lifestyles.

Index Terms— lifelogging datasets, lifestyle analysis,activity analysis, healthy degree, latent fluents

1. INTRODUCTION

A healthy lifestyle is the key to better health and happiness, and has a considerable effect on the quality of life and disease prevention [1]. With the rapid development of wearable devices, visual lifelogging has attracted increasing research attention. Visual lifelogging consists of acquiring images that capture daily experiences of users who wear a camera over a long period of time and automatically analyzing the activities based on the captured egocentric data [2]. Most of current visual lifelogging research focuses on daily activity recognition, such as cooking and working. In this paper, we analyze lifestyle of a user based on visual lifelogging, which can help the user to establish a healthy lifestyle.

To analyze a user's lifestyle via visual lifelogging, data captured by a wearable device should cover the entire day and be captured at intervals shorter than the duration of any activity of interest. In theory, continuously capturing video for an entire day is optimal; however, currently available wearable devices cannot capture videos for a whole day due to storage and battery limitations. Currently, available lifelogging/egocentric datasets can be classified into two categories:

1) Datasets with images taken every tens of seconds to record the daily life of a user for a long period of time, such as AIHS [3], ImageCLEFlifelog2020 [4] and EDBU [5]. As the time interval between images is fairly large, some details of daily activities are missing and some activities with short durations cannot be captured. For example, Figure 1(a) shows the images captured every 30 seconds and images captured every 3 seconds.

2) Datasets with video clips to record certain activities, with each video clip typically corresponding to a single activity, such as the Extended GTEA Gaze+ [6], EPIC-Kitchens Dataset[7], UT Ego [8] and EgoSeg [9]. These datasets are not suitable for lifestyle analysis as the video clips do not cover a long period of time, such as a complete day. For example, the top row of Figure 1(b) shows two videos in GTEA Gaze+ and UT Ego.

Current lifelogging datasets are either missing details about activities (because of the large time interval between the captured images) or do not cover a long period of time (video clips of different activities are captured separately) and are thus not suitable for lifestyle analysis. Based on the above observations, we build a visual lifelogging dataset for lifestyle analysis (VLDLA) with images captured with a wearable camera every three seconds from 8:00 am to 6:00 pm for seven days.

Based on our dataset, we propose three latent fluents to measure the healthiness of a lifestyle. The three latent fluents are fatigue, to indicate whether the user rests regularly; thirst, to indicate whether the user drinks regularly; and hunger, to indicate whether the user eats regularly. The three latent fluents change over time; for example, a working user will become increasingly tired as time goes on.

Activities of a user must be recognized first to analyze the three latent fluents. Many methods have been proposed for egocentric activity analysis. Swathikiran Sudhakaran et al. [10] propose an end-to-end two-stream long short-term attention(LSTA) network. Evangelos Kazakos et al. [11] propose an end-to-end trainable audio-visual temporal binding network(TBN) for ego-

centric action recognition. Kyle Min et al. [12] combine spatiotemporal attention with human gaze for egocentric activity recognition.Minjie Cai et al. [13] propose a Bayesian CNN to transfer existing hand segmentation algorithm to new unlabeled datasets.

In this paper, we classify activities in each frame by combining scene context, object context and temporal information. The proposed method is based on the following observations. (1). Specific activities always occur in certain scenes, and scene context can provide important information for activity classification. (2). Objects are important for many daily activities, and object context can also help to classify activities. (3). There are temporal correlations between frames of each activity and temporal constraints between different activities, which can be used for activity classification.

The main contributions of this paper can be summarized as follows: (1). This study is the first attempt to analyze a user's lifestyle via visual lifelogging, and we build the VLDLA with images captured every three seconds from 8:00 am to 6:00 pm for seven days. The VLDLA is suitable for lifestyle analysis as it covers all the activities of each day in sufficient detail. (2). We analyze the lifestyle based on three latent fluents, namely, fatigue, hunger and thirst, which reflect the user's condition. (3). We classify the daily activities of the user based on the scene context, object context, temporal correlations between frames in each activity, and temporal constraints between different activities simultaneously.

2. DATASET

We build the VLDLA by a FrontRow wearable camera. FrontRow is a portable wearable lifelogging device that can be worn on the chest for 147.5 ° wide-angle shooting.

The VLDLA contains 84,000 images with a resolution of 1920×1080 captured every three seconds from 8:00 am to 6:00 pm for seven days. There are twelve distinct activities: using computer, reading, using phone, attending class, walking, resting, exercising-outdoor, exercising-indoor, shopping, eating, drinking and social.

For lifestyle analysis, some activities, such as using computer, using phone and reading, are not substantially different as they are all sedentary activities. Therefore, we categorize the twelve activities into 5 groups: Sedentary Work, Motion, Shopping, Food and Social. Intuitively, for a healthy lifestyle, sedentary activities should not last excessively long periods of time, and food-related activities such as eating and drinking should occur regularly throughout the day. As shown in Figure 4, sedentary activities, food and motion are more well distributed for a healthy lifestyle then an unhealthy lifestyle.

Privacy is a major concern for lifelogging data. We use the face detector in [14] and screen detector in [15] to detect the faces and screens in each frame and smooth

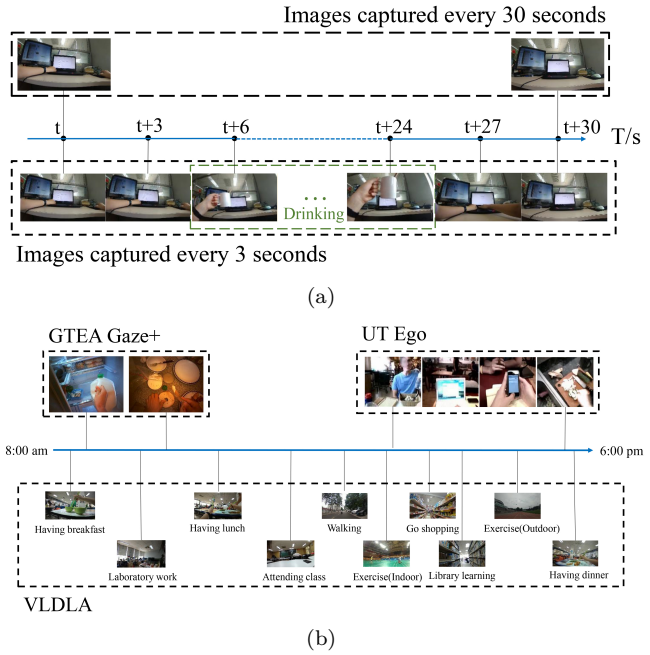

Fig. 1. (a). Our dataset contains more details about daily activities, such as drinking, that are missing from datasets with images captured every 30 seconds. (b). In contrast to datasets with separate videos, our dataset covers the activities of a whole day and is suitable for lifestyle analysis.

the detected faces and screens with Gaussian filters using a sufficiently large variance[16] to protect the privacy of the people captured in our dataset. As shown in Figure 2.

Each frame in VLDLA is annotated as one of the twelve activities by the user who records data. Lifestyle is usually defined based on a long period of time. Here, we analyze the lifestyle each day and assign a score from 0 to 1 to indicate whether the lifestyle is healthy: 1 indicates a perfectly healthy lifestyle, and 0 indicates a absolutely unhealthy lifestyle. We ask 10 participants, including 4 females and 6 males, from a local university, whose ages range from 18 to 25 years, to score the lifestyle for each day by showing them the script generated by labels of each frame.

3. METHOD

3.1. Lifestyle Analysis

We define three latent fluents to compute the lifestyle score: hunger, thirst, and fatigue. We analyze the lifestyle according to commonly accepted assumptions about a healthy lifestyle, that is, one should take a break after sedentary work and one should eat and drink regularly. We compute a score for each latent fluent in each frame, and the overall lifestyle score is computed as

$$s_j^{lifestyle} = 1 - \frac{1}{3N_j}\sum_{i=1}^{N_j}(h_i + t_i + f_i) \quad (1)$$

Fig. 2. Sample frames from our VLDLA dataset. Note that the faces and screens in the frames are blurred for privacy protection.

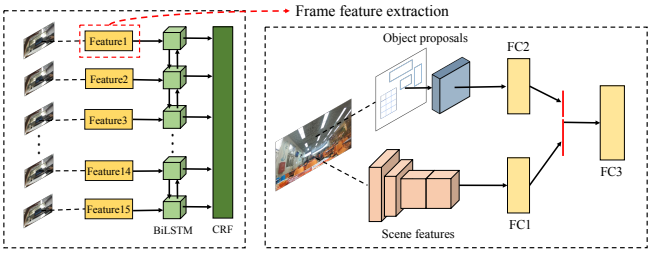

Fig. 3. Activity recognition based on scene features, object features and temporal constraints. Scene features from ResNet50 and object proposals from Mask R-CNN are fused as frame feature, and is fed into a BiLSTM-CRF for activity recognition.

where N_j is the total number of frames for the jth day and $3N_j$ is used to normalize the lifestyle score to $(0, 1)$. We use 1 minus the average score of the three latent fluents to make the lifestyle score accord with the convention that 1 corresponds to a healthy lifestyle and 0 corresponds to an unhealthy lifestyle. h_i, t_i and f_i are the scores for the three latent fluents in frame i.

h_i is computed as

$$h_i = \begin{cases} 0 & if \quad c_i = eating \\ \dfrac{1}{1 + e^{-(\frac{i - i_{eating}}{1200} - \alpha_h)}} & else \end{cases} \tag{2}$$

where c_i is the activity occurring in frame i, i_{eating} is the frame index of the last eating activity, 1200 is the number of frames captured in an hour, and α_h is a hyperparameter that is set to 5 since a person will typically become hungry approximately 5 hours after eating. h_i will be 5 hours after eating and will increase further as time passes.

t_i and f_i are computed similarly to h_i, and α_t and α_f are defined similarly to α_h and are set to 2 and 1, respectively.

3.2. Activity Recognition

To compute the scores for the latent fluents, the user activities must be recognized first, as many activities can change the latent fluents. We recognize the activity in each frame based on the scene features, object features and temporal information.

Most of these twelve activities have their own specific occurrence scenarios. For example, eating generally occurs in a cafeteria, restaurant, food_court, etc. We use the ResNet50 [17] trained on Places365 [18] to extract scene features for each frame.

Activities are also generally associated with representative objects, such as saucers and bowls for eating. We use the Mask R-CNN [19] trained on COCO to extract object proposals in each frame as object features.

As shown in Figure 3, we replace the last original fully connected layer of ResNet 50 with a new fully connected layer (FC1) with a dimension of 500 and treat the 500-dimensional output of the FC1 layer as the scene feature. We add a fully connected layer (FC2) with a dimension of 400 after the object proposals from the RoiAlign layer of Mask R-CNN and treat the 400-dimensional output of the FC2 layer as the object feature. The scene feature and object feature are concatenated and passed to a fully connected layer (FC3) with a dimension of 500, and the 500-dimensional output of the FC3 layer is used as the combined features.

The combined features are fed into a BiLSTM-CRF[20] for activity recognition. The BiLSTM can learn the temporal correlations between the combined features, and the linear-chain CRF can impose temporal constraints between activities, for example, resting usually follows sedentary work and a person will not eat more than once within a short period of time. For the BiLSTM, we use the same implementation as [20].

We use batches of $n(15)$ consecutive frames as the input of BiLSTM, and classification result from CRF is $y = (y_1, y_2, y_3, ..., y_n)$. Score of y can be computed as

$$s(X, y) = \sum_{i=0}^{n} A_{y_i, y_{i+1}} + \sum_{i=1}^{n} P_{i, y_i} \tag{3}$$

where P is the $n * k$ matrix of scores output by the BiLSTM, k is the number of activity classes, and $P_{i,j}$ corresponds to the score of the jth label of the ith frame in a video. A_{y_i, y_j} represents the transition score between label y_i and label y_j. For more details, please refer to[20].

4. EXPERIMENT

4.1. Activity Recognition

We split the VLDLA dataset into a training set and testing set.The data from the first, second, third, and sixth days are chosen as the training data, and the remaining data are used for testing.

We compare our method with InceptionV3+RF+LSTM in [21], which achieves the best performance among the many methods compared in [21]. The comparison of the accuracy, macro precision, macro recall and macro

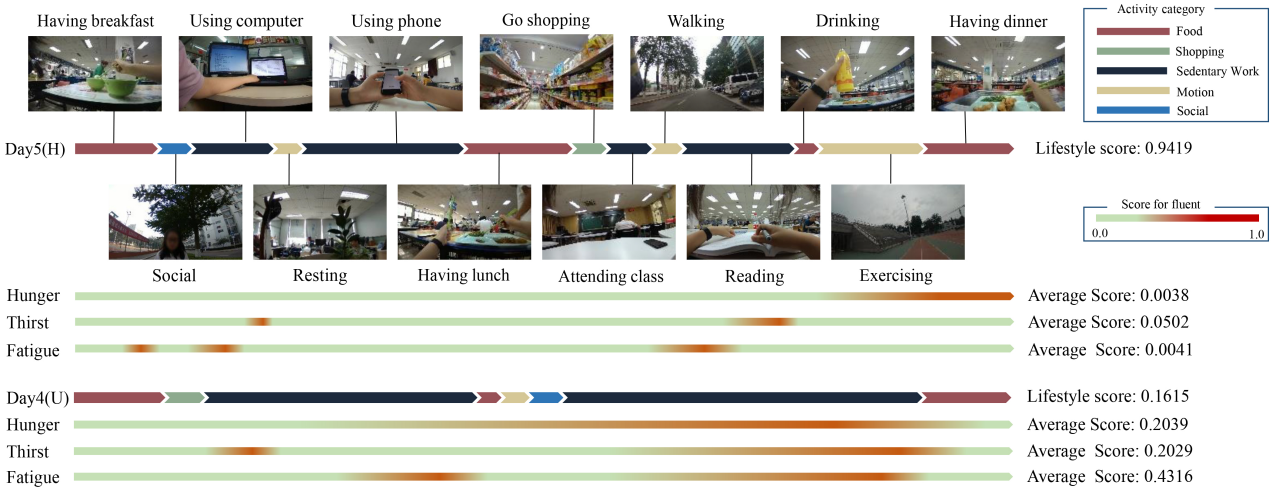

Fig. 4. Demonstration of scores for three latent fluents and lifestyle. For Day 5, the user has a healthy lifestyle: sedentary work, food and motion are well-distributed, the scores for hunger, thirst and fatigue are low and the score for lifestyle is high. For Day 4, the user participates in excessive sedentary work, and the score for lifestyle is low.

Table 1. Comparison results of the F1-score for each activity category

Method	Social	Using computer	Reading	Using phone	Attending class	Walking	Resting	Exercising (outdoor)	Exercising (indoor)	Shopping	Eating	Drinking
InceptionV3+RF+LSTM [21]	0.5356	0.7039	0.6045	0.0587	0.7501	0.6717	0.5240	0.5198	0.0267	0.3810	0.6591	0.1032
Our method	0.4631	0.9064	0.8300	0.4548	0.6488	0.8954	0.9757	0.9593	0.9628	0.6990	0.9564	0.3858

Table 2. Overall comparison of activity recognition

Method	Accuracy	Macro Precision	Macro Recall	Macro F1-score
InceptionV3+RF+LSTM [21]	0.5669	0.6311	0.4869	0.4615
Our method	0.8557	0.7695	0.8028	0.7615

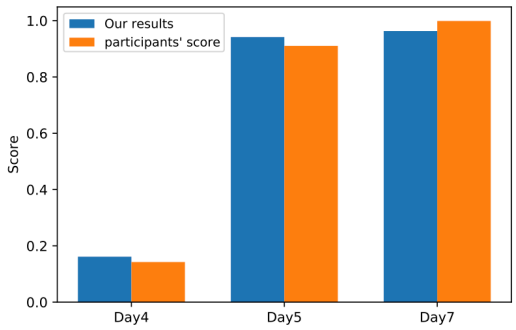

Fig. 5. Experimental results for the lifestyle analysis.

F1-score is shown in Table 2, and the comparison results of the macro F1-score for each activity category are shown in Table 1. Table 2 and Table 1 illustrate that our method achieves better performance on most categories because the BiLSTM-CRF can simultaneously impose temporal correlations between the frames in each activity and temporal constraints between different activities.

4.2. Lifestyle Analysis

Based on the activity recognition results, we compute scores for the three latent fluents and for the overall lifestyle. Figure 4 shows the demonstration of scores for the three latent fluents and lifestyle. For Day 5, the user has a healthy lifestyle: sedentary work, food and motion are well-distributed, scores for hunger, thirst and fatigue are low and the score for lifestyle is high. For Day 4, the user participates in excessive sedentary work, and the score for lifestyle is low. We compare the computed scores with the participants' scores as shown in Figure 5, where the lifestyle of Day 4 is unhealthy, and the lifestyle of Day 5 and Day 7 is healthy. The computed score for lifestyle accords with the participants' score, which confirms the rationality and effectiveness of our method.

5. CONCLUSION

In this work, we build a new visual lifelogging dataset for lifestyle analysis (VLDLA) that contains images taken every three seconds for seven days. The VLDLA covers a long period of time with images captures at short time intervals and is suitable for lifestyle analysis. Based on VLDLA, we propose a method for lifestyle analysis based on three latent fluents and recognition of daily activities.

6. ACKNOWLEDGMENTS

This work was supported by the Natural Science Foundation of China (NSFC) under Grant No.61972038.

7. REFERENCES

[1] Yanping Li, Pan An, Dong D. Wang, Xiaoran Liu, Klodian Dhana, Oscar H. Franco, Stephen Kaptoge, Emanuele Di Angelantonio, Meir Stampfer, and Walter C. Willett, "Impact of healthy lifestyle factors on life expectancies in the us population," Circulation, p. CIRCULATIONAHA.117.032047, 2018.

[2] Marc Bolanos, Mariella Dimiccoli, and Petia Radeva, "Toward storytelling from visual lifelogging: An overview," IEEE Transactions on Human-Machine Systems, pp. 77–90, 2017.

[3] Nebojsa Jojic, Alessandro Perina, and Vittorio Murino, "Structural epitome: a way to summarize one's visual experience," in Advances in neural information processing systems, 2010, pp. 1027–1035.

[4] ImageCLEFlifelog2020, ," https://www.imageclef.org/2020/lifelog, 2020, Accessed January 10, 2020.

[5] Marc Bolaños and Petia Radeva, "Ego-object discovery," arXiv preprint arXiv:1504.01639, 2015.

[6] Yin Li, Miao Liu, and James M Rehg, "In the eye of beholder: Joint learning of gaze and actions in first person video," in ECCV, 2018, pp. 619–635.

[7] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, et al., "Scaling egocentric vision: The epic-kitchens dataset," in ECCV, 2018, pp. 720–736.

[8] Zheng Lu and Kristen Grauman, "Story-driven summarization for egocentric video," in CVPR, 2013, pp. 2714–2721.

[9] Yair Poleg, Chetan Arora, and Shmuel Peleg, "Temporal segmentation of egocentric videos," in CVPR, 2014, pp. 2537–2544.

[10] Swathikiran Sudhakaran, Sergio Escalera, and Oswald Lanz, "Lsta: Long short-term attention for egocentric action recognition," in CVPR, 2019.

[11] Evangelos Kazakos, Arsha Nagrani, Andrew Zisserman, and Dima Damen, "Epic-fusion: Audio-visual temporal binding for egocentric action recognition," in ICCV, 2019.

[12] K. Min and J. J. Corso, "Integrating human gaze into attention for egocentric activity recognition," 2020.

[13] M Cai, F. Lu, and Y. Sato, "Generalizing hand segmentation in egocentric videos with uncertainty-guided model adaptation," in 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020.

[14] Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, and Qiao Yu, "Joint face detection and alignment using multitask cascaded convolutional networks," IEEE Signal Processing Letters, pp. 1499–1503, 2016.

[15] Mohammed Korayem, Robert Templeman, Dennis Chen, David Crandall, and Apu Kapadia, "Enhancing lifelogging privacy by detecting screens," in Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems. ACM, 2016, pp. 4309–4314.

[16] Slobodan Ribaric, Aladdin Ariyaeeinia, and Nikola Pavesic, "De-identification for privacy protection in multimedia content: A survey," Signal Processing: Image Communication, pp. 131–151, 2016.

[17] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in CVPR, 2016, pp. 770–778.

[18] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba, "Places: A 10 million image database for scene recognition," IEEE transactions on pattern analysis and machine intelligence, pp. 1452–1464, 2018.

[19] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick, "Mask r-cnn," in ICCV, 2017, pp. 2961–2969.

[20] Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer, "Neural architectures for named entity recognition," arXiv preprint arXiv:1603.01360, 2016.

[21] Alejandro Cartas, Juan Marín, Petia Radeva, and Mariella Dimiccoli, "Batch-based activity recognition from egocentric photo-streams revisited," Pattern Analysis and Applications, pp. 953–965, 2018.