Leveraging Content and Context in Understanding Activities of Daily Living

Minh-Son Dao[1,2], Asem Kasem[1], and Mohamed Saleem Haja Nazmudeen[1]

[1] Universiti Teknologi Brunei, Brunei
{minh.son,asem.kasem,mohamed.saleem}@utb.edu.bn
www.utb.edu.bn
[2] University of Information Technology, Vietnam
sondm@uit.edu.vn
www.uit.edu.vn

Abstract. This paper introduces a content-context-based method to automatically provide a summarization of lifelog data based on selected concepts of Activities of Daily Living (ADL). The main idea of the proposed method is to create a so-called (1) Daily-Normal Environment Panorama (DNEP) image, and a (2) Daily-Abnormal Environment (DAE) Taxonomy. The former is used to detect events that happen in known environments such as in a house, in an office and on the way from a parking lot to an office. The latter is used to detect events whose concepts can be detected by a pre-defined taxonomy such as in a restaurant, in a church, and on a street. The proposed method is evaluated by using the data and evaluation tool offered by organizers of imageCLEFlifelog2018 - subtask Activities of Daily Living understanding (ADLT). The experiments show that the proposed method works better than methods proposed by other participants of the imageCLEFlifelog2018.

Keywords: lifelog data analysis · image alignment and stitching · semantic taxonomy · heterogeneous data segmentation

1 Introduction

Lifelogging is an interesting topic that can help to understand the daily living activities and to recall moments of interest that happened in the past. The scope of applications that utilize lifelog data extends to human-supported scenarios such as personal healthcare and personal assistants [1]. Since lifelog data is a time-series data, events detection in lifelog data can be considered as scene detection in video data where data segmentation plays an essential role [2]. Therefore, the task of segmentation in lifelog data analysis is crucial, and recently many works have focused on lifelog data segmentation [3][4][5][6]. Most of these methods utilized visual features only for video segmentation. However, lifelog data contains not only visual information, but also heterogeneous data such as textual data (e.g. tags, comments), geo data (e.g. GPS, places name), and physiological data (e.g. heartbeat, skin temperature). Hence, it is essential to have a method that

can analyze such heterogeneous and big data to extract the information people may need.

The imageCLEFlifelog2018 (the organizer hereafter) [7][8] organizes a challenge to encourage participants to propose methods to automatically analyze such data towards categorizing, summarizing, and retrieving information of interest. We have joined the challenge with the subtask "Activities of Daily Living understanding" (ADLT), and this paper reports our work and discusses its evaluation on this subtask.

2 Methodology

In this section, we introduce what we call DNEP image and DAE taxonomy and how to leverage them to detect events in lifelog data. The idea behind these image and taxonomy concepts is based on "Content without Context is meaningless" [9]. It means that we use not only the content of lifelog data to understand an event, but also the context where that event happened. This is expected to improve event detection as well as decrease the missing/redundant information of events boundaries.

2.1 Daily-Abnormal Environment Taxonomy (Contexts and Activities)

Most of the events that happen daily have their own unspoken/spoken rules by which we can build a suitable taxonomy. For example, when visiting a church the environment must contain salient and typical symbols of a church such as the cross and Saint statutes, while the activities could be: slowly walking, quietly sitting or kneeling. Therefore, based on the events concept we can build a taxonomy for the specific event to further detect that event in lifelog data. Another example is socialising in a restaurant. The visual taxonomy of a meal table, especially a menu and a counter (with or without a queue), can be integrated with GPS and/or the restaurants name tagged by people to distinguish whether the event happened in a restaurant or in a relatives house.

In fact, each daily activity can be determined when knowing the environment where such an activity happens. These environments again can be determined by "scene recognition" and "visual concepts detection" tasks. The former names a place and the latter labels all objects appeared inside the place.

The work carried out by Zhou et al. [13] is a good example of scene recognition. In this work, the *scene hierarchy* defined by the authors has two levels, and each scene is located to a suitable slot in this *scene hierarchy*. For example, the *conference room* scene is located at *(level 1: indoor → level 2: workplace (office building, factory, lab, etc.))*[3]. *The organizer* also offered the *scene ontology* described in *NTCIR-13 Lifelog Ontology* [8]. This ontology gives the summary of scenes appeared in the dataset. By integrating the *scene ontology* and *scene*

[3] *http://places2.csail.mit.edu/download.html*

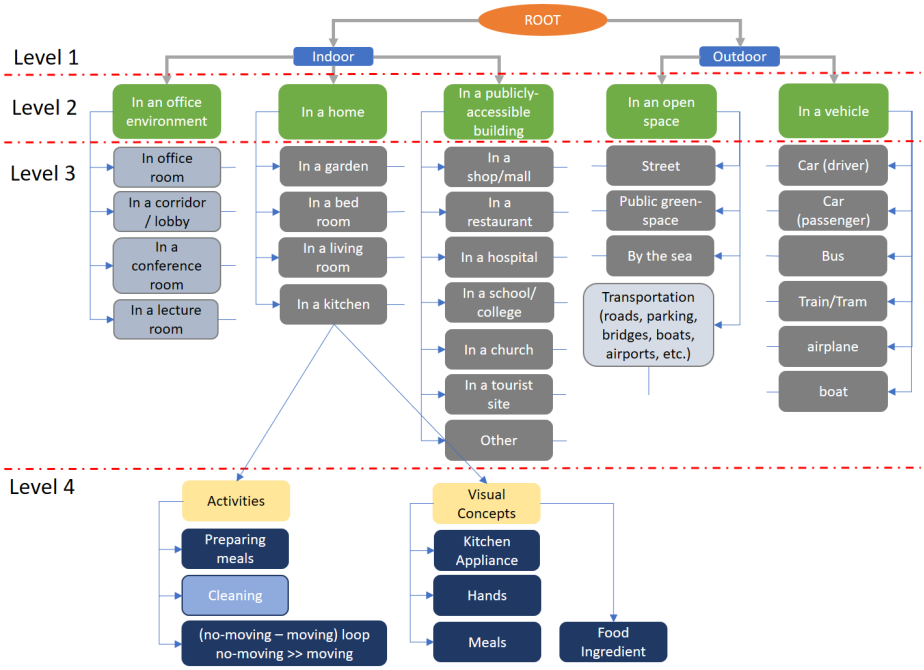

Fig. 1. An example of DAE taxonomy of topic 3 "preparing meals, home"

hierarchy, we build DAE taxonomy from the root to level 3, as illustrated in Fig. 1.

When successfully locating scene A (in level 3), the next question is "what kind of activities normally happens and which visual concepts always appear in the scene A?". Such a question leads to the need of building level 4 of DAE taxonomy. In this level, each level-3 scene has two same-level categories *activities* and *visual concepts*. The former is built based on the *Activities/Facets of Life activity* defined in *NTCIR-13 Lifelog Ontology*; and the latter is constructed by utilizing the *visual concepts* and *food-logs* and *drink-logs* described in [8]. For example, *in a kitchen*, *preparing a meal* often happens and definitely there must be *kitchen appliance* and *foods*, illustrate in Fig. 1.

Algorithm 1: DAE taxonomy building

1. Repeat Alg.1-step.1 to the ground-truth data to generate the training data.
2. Use the context, physical activities tags, physiological data, and image/visual concepts provided by *the organizer* and *scene hierarchy* offered by [13] to build the DAE taxonomy. In our case, the physical activities tags and physiological data are utilized to determine two categories (1) moving (2) no-moving.
3. Build a lookup location table (DAE-LT) for determining the geofencing of the event using GPS and places name tags.

The moving of people inside a scene also can give rich semantic cues for understanding their activities. For example, the loop of long standing (no-moving) and short walking (moving) inside a kitchen can differ from the loop of long sitting (no-moving) and short walking/turning (moving) inside an office. These information can be captured by using *biometrics:(grs, steps)* entry in the metadata of dataset with the *human activities recognizer* developed in [15].

The Algorithm 1 summarizes the content of this subsection.

2.2 Daily-Normal Environment Panorama Image (Visual Background)

One of the characteristics of lifelog data is the repeated routine. People usually have at least one place to visit almost every day and the environment of this place rarely changes, such as at home, a relatives house, an office, a favorite restaurant, and a familiar supermarket. Therefore, if we can accumulate all images captured from those places, we can build a panorama image. Consequently, if we can successfully project a lifelog image onto a panorama image (e.g. using image alignment, object detection, image segmentation) with a known concept, we can assign the right event label for that image and further detect the boundary of that event.

The Image Alignment and Stitching have been researched for over a decade now [10]. Two popular approaches are applied to align and stitch images (1) features-based, and (2) direct (or global) methods. While the former uses images features (e.g. points, edges), the latter utilizes the whole image to estimate the transformation between images. We utilized two methods introduced by Meneghetti et al. [11] and Poleg and Peleg [12] to create our DNEP images. The former can deal with the sparsely structured environment where not enough distinct features can be detected such as in the case of uniform walls, floors and ceilings in indoor scenarios, and sky and sea in outdoor scenarios. The latter can deal with non-overlapping images issue that happens due to non-continuous recording if lifelog data.

After creating a DNEP image, this image will be located into the DAE taxonomy by using *scene recognition* tools [13]. The Fig. 2 illustrates the DNEP image of *in a living room* scene that created by aligning and stitching all developing data containing living room images.

The Algorithm 2 denotes the way we create DNEP image.

Algorithm 2: DNEP image building

1. Use the ground-truth data to build the training data of a given event.
2. Assign the suitable hierarchical context (e.g. indoor $\rightarrow$ in a house $\rightarrow$ in a kitchen, indoor $\rightarrow$ in a house $\rightarrow$ in a living room, indoor $\rightarrow$ in a working place $\rightarrow$ in an office) to the training data.
3. Build a lookup location table (DNEP-LT) for determining the geofencing of the event using GPS and places name tags.
4. Utilize algorithms introduced in [11] and [12] to build events DNEP image.

Fig. 2. In a living room

Fig. 3. In an office 1

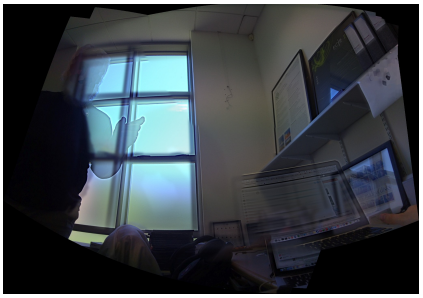

Fig. 4. In an office 2

In fact, DNEP image is a set of panorama images created by aligning and stitching all images contained in the training data. In the beginning, there could be several panorama images due to not having enough images to reflect the surrounding environment. Nevertheless, these panorama images will be merged if more images are added to the training dataset. This process can be done either by manually adding more ground-truth data or by automatically importing the images generated by Algorithm 3.

The Fig. 3 and 4 illustrate one example of DNEP image. In this case, there are two DNEP images that express the same environment *office*. Due to lack of suitable images they have temporally not yet merged together. Nevertheless, there is a *ghost* laptop-monitor that could be a good cue for merging two DNEP images. In this paper, this problem has not yet solved and is served for the future work.

2.3 Activities of Daily Living Understanding Using DNEP image and DAE taxonomy

The idea of our content-context-based segmentation algorithm is quite simple. We divide the environment into two distinguish categories: (1) *daily-normal environment*, and (2) *occasionally-abnormal environment*. For example, the *in a*

home and *in an office* environments are assigned to the former, and the *in a publicly-accessible building* and *in an open space* are classified to the latter one.

With the *daily-normal environment* category, we create DNEP images and treat them as the visual background of activities, and access DAE taxonomy by top-bottom direction. Giving an image a, we simply use the object detection or template matching to justify whether this image belongs to the DNEP image of the required environment A. We than use DAE taxonomy to check how many visual concepts and activities of the current scene A the image a can satisfy to get the conclusion of what activity it is.

With the *occasionally-abnormal environment*, we apply DAE taxonomy in the bottom-up manner. It means that, first we try to extract as many visual concepts as possible from a given image to fulfil *visual concepts* of level 4. Then, we recognize the scene of the given image to match level 3 to the root. Other tasks such as location name extraction and activities detection are carried out parallel to fill the *activities* of level 4. Finally, we check whether the new taxonomy (just created) is matched with the DAE taxonomy of the required task.

The Algorithm 3 abstractly describes how we can carry out the content-context-base segmentation to meet the requirement of the challenges.

Algorithm 3: Content-Context-based Segmentation

1. Prepare DNEP images and DAE taxonomies based on the quest.
2. Pick an event. Based on the events concept, it could have a DNEP image and/or DAE taxonomy that can be used for segmentation.
3. IF using DNEP image
   (a) create the CANDIDATE-A buffer as $\{(image - id, event/non - eventlabel)\}$
   (b) load the related DNEP image.
   (c) with each lifelog image that falls in the temporal frame defined by the event (e.g. watching TV before 7am) determine whether this image belongs to the DNEP image by applying object detection (with occlusion option).
      i. if successful, assign the event label to this image, and add to the buffer CANDIDATE-A.
      ii. if this image does not appear in the DNEP image but its location is still in the DNEP images geofencing, then use the algorithm in [12] to align and stitch this image to the DNEP image. Next, assign the event label to this image, and add it to the buffer CANDIDATE-A.
      iii. if this image and its location do not belong the DNEP image, then it is assigned non-event label and add to the buffer CANDIDATE-A.
   (d) repeat (c) until all temporal frame is scanned.
   (e) merge all consecutive *(image-id, event label)* of CANDIDATE-A if a certain number of non-event labels lay between two event clusters. In our case, this number is set to be less than 1/10 of total time when merging them together.
   (f) those images that are assigned as *non-event-label* before merging and as *event-label* after merging will be sent to the CANDIDATE-A set.

Further, they will be manually confirmed and automatically aligned and stitched to the relative DNEP image. This will help to decrease the number of panorama images as well as to increase the coverage of the DNEP image.

4. IF using DAE taxonomy
   (a) create the CANDIDATE-B buffer as $\{(image-id, event/non-eventlabel)\}$
   (b) load the related DAE taxonomy.
   (c) with each lifelog image that falls in the temporal frame defined by the event,
      i. detect all objects and scenes defined in the DAE taxonomy. **NOTE**: In our case, we used *visual concepts* that contain both objects and scenes names, provided by *the organizer* for this task.
      ii. extract information of locations *(GPS, place's name)*, and physical activities *(action's names, moving/no-moving)*.
      iii. check whether this information satisfies the DAE taxonomy. If yes, add it to the CANDIDATE-B buffer as *event-label*. If not, add it as *non-event-label*.
   (d) merge all consecutive *(image-id, event label)* of CANDIDATE-B similarly to 3(e) above.
5. IF both DNEP images and DAE taxonomies are used
   (a) since we treat a DAE taxonomy as a foreground and a DNEP image as a background, we merge the CANDIDATE-A and CANDIDATE-B so that CANDIDATE-A should cover CANDIDATE-B.

3 Experiments

The data and metrics offered by imageCLEFlifelog2018 - subtask Activities of Daily Living understanding (ADLT) are utilized to evaluate the proposed method. Ten events with given concepts are required to be detected; each detected event must be reported in the form of a triplex (topic-id, number-of-times, number-of-minutes) where topic-id is the number of the queried topic, numberof-times reports how many times the event occurred, and number-of-minutes tells us for how long (in minutes) the event lasted. Equation (1) is the metric used to evaluate our results.

$$ADL_{score} = \frac{1}{2}(max(0, 1 - \frac{|n - n_{gt}|}{n_{gt}}) + max(0, 1 - \frac{|m - m_{gt}|}{m_{gt}})) \qquad (1)$$

where (n, n_{gt}) and (m, m_{gt}) are the *(submitted value, ground-truth value)* for how many times the events occurred, and for how long (in minutes) the events lasted, respectively.

Based on the testset provided by *the organizer*, we assigned categories of DNEP image and DAE taxonomy to the ten required queries as denoted in Table 1

We used the same parameters of the methods we have utilized. Table 2 reports the results of participants in this subtask.

Table 1. Dividing ten required queries into DNEP image and DAE taxonomy

Query ID	Query (ADL, context)	Category
1	(Drinking coffee, in an Office)	DNEP image, DAE taxonomy
2	(Shopping, outside Office)	DAE taxonomy
3	(Preparing meals, Home)	DNEP image, DAE taxonomy
4	(Watching TV, Home)	DNEP image, DAE taxonomy
5	(Listening/Watching Presentations, at Work)	DAE taxonomy
6	(Using mobile device, In a vehicle)	DAE taxonomy
7	(Not using computers, In an office)	DNEP image, DAE taxonomy
8	(Walking, on the street)	DAE taxonomy
9	(, In a church)	DAE taxonomy
10	(Socialising/Eating/Drinking, In a restaurant)	DAE taxonomy

Table 2. Activites of Daily Living Understanding Competitive Results

Group name	Percentage Dissimilarity	Rank Dissimilarity
CIE@UTB (our group)	**0.556**	**1**
NLP-Lab	0.479	2
HCMUS	0.059	3

In fact, the proposed method works case-by-case since it heavily depends on the content and context of a given event. The training phase is vital and needs manual intervention to build a suitable taxonomy depending on given activities and contexts. While the DNEP image can be generally generated without or with little manual support, the DAE taxonomy is established using peoples knowledge about the event and how many objects the system can detect and recognize from images. Thus, if the person lacks events knowledge to build the DAE taxonomy, the result of event segmentation could degrade.

For some events, both DNEP image and DAE taxonomy did not work well, e.g. in the case of *"Find how many times and how long the user is having coffee in the office. Having coffee at the bars at the workplace is not considered."* (topic 1 - subtask ADLT). The DNEP image successfully detects an office scene, and the DAE taxonomy can recognize a coffee cup on a table. Nevertheless, it is hard to find a suitable hint to know exactly when the person drinks the coffee. Although that person already tagged the time for drinking coffee, visual and other cues say nothing about that activity. In this case, our method almost failed to detect the right boundary of this event

4 Conclusions

In this paper, we introduce a content-context-based method to automatically detect events with given concepts from lifelog data. Data and metrics offered by imageCLEFlifelog2018 are used to evaluate the proposed method. The daily-normal environment panorama image and the daily-abnormal environment taxonomy are created to detect events. The events content (e.g. visual, textual,

physiological, and GPS features) and context (e.g. concepts and taxonomy) are carefully taken into account to create DNEP image and DAE taxonomy as well as to detect events. Both DNEP image and DAE taxonomy have the ability to evolve themselves along the lifelog data time. It means that the more data gets recorded, the larger the scope of events DNEP image and DAE taxonomy can cover. In future, post-processing to polish events boundaries will be investigated. Moreover, fusion of features [14] will be evaluated to seek better evaluation. Currently, we only use features provided by the organizer. These features are somehow not enough for building a strong DAE taxonomy as well as successfully projecting an image into DNEP images. Consequently, we will develop our own features extractors to fulfil our requirements.

References

1. Gurrin, C., Smeaton, A.F., Doherty, A.R.: LifeLogging: Personal Big Data. Journal of Foundations and Trend ^{R} in Information Retrieval. **8**(1), 1-125 (2014)
2. Del Fabro, M., Böszörmeny, L.: State-of-the-art and future challenges in video scene detection: a survey. Journal of Multimedia Systems. **19**5, 427–454 (2013)
3. Doherty, A.R., Smeaton, A.F., Lee, K., Ellis, D.P.W.:. Multimodal segmentation of lifelog data. In: Procs. Large Scale Semantic Access to Content (Text, Image, Video, and Sound) (RIAO '07), pp. 21–38 Paris, France (2007)
4. Dimiccoli, M. et al.: SR-clustering: Semantic regularized clustering for egocentric photo streams segmentation. Journal of Computer Vision and Image Understanding. **155**, 55–69 (2017)
5. Gupta, R., Gurrin, C.: Approaches for Event Segmentation of Visual Lifelog Data. In: MultiMedia Modeling, pp. 581–593 (2018). https://doi.org/10.1007/9783319736037-47
6. Furnari, A., Battiato, S., Farinella, G.M.: Personal-Location-Based Temporal Segmentation of Egocentric Videos for Lifelogging Applications. Journal of Visual Communication and Image Representation. **12**, 1–12 (2018)
7. Ionescu, B., Muller, H., Villegas, M., de Herrera, A.G.S., Eickhoff, C., Andrearczyk, V., Cid, Y.D., Liauchuk, V., Kovalev, V., Hasan, S.A., Ling, Y., Farri, O., Liu, J., Lungren, M., Dang-Nguyen, D.T., Piras, L., Riegler, M., Zhou, L.m Lux, M., Gurrin, C.: Overview of ImageCLEF 2018: Challenges, Datasets and Evaluation. In: Experimental IR Meets Multilinguality, Multimodality, and Interaction, CLEF 2018, Avignon, France (2018)
8. Dang-Nguyen, D.T., Piras, L., Riegler, M., Zhou, L., Lux, M., Gurrin, C.: Overview of ImageCLEFlifelog 2018: Daily Living Understanding and Lifelog Moment Retrieval. In: Procs. CEUR Workshop, CLEF2018 Working Notes, Avignon, France (2018)
9. Jain, R., Sinha, P.: Content without Context is Meaningless. In: Procs. ACM MM 2010, pp. 1–10. ACM, Firenze, Italy (2010)
10. Szeliski, R.: Image Alignment and Stitching: A Tutorial. Journal of Foundations and Trend ^{R} in Computer Graphics and Vision. **2**(1), 1–104 (2006)
11. Menegetti, G., Danelljan, M., Felsberg, M., Nordberg, K.: Image Alignment for Panorama Stitching in Sparsely Structured Environments. In: R.R. Paulsen and K.S. Pedersend (Eds) SCIA 2015, LNCS 9127, pp. 428–439 (2015). https://doi.org/10.1007/9783319196657-36

12. Poleg, Y., Peleg, S.: Alignment and Mosaicing of Non-Overlapping Images. In: Procs. IEEE Int. Conf. on Computational Photography (ICCP) (2012)
13. Zhou, B., Lapedriza, A., Khosla, A., Oliva, A., Torralba, A.: Places: A 10 million Image Database for Scene Recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1–14 (2017)
14. Dao, M.S., Pham, Q.N.M., Kasem, A., Nazmudeen, M.S.: A Context-Aware Late-Fusion Approach for Disaster Image Retrieval from Social Media. In: ACM ICMR, Yokoham, Japan (2018)
15. Dao, M.S., Dang-Nguyen, D.T., Riegler, M., Gurrin, C.: Smart Lifelogging: Recognizing Human Activities using PHASOR. ICPRAM 2017 (2017)