# Repeated Event Discovery from Image Sequences by using Segmental Dynamic Time Warping: experiment at the NTCIR-12 Lifelog task

Kosuke Yamauchi and Tomoyosi Akiba
Toyohashi University of Technology
{yamauchi | akiba}@nlp.cs.tut.ac.jp

## ABSTRACT

In this paper, we report on examining in applying of Spoken Term Discovery (STD) to lifelog images. STD is an approach to discover words which repeated in multiple speeches. We used an approach based on Dynamic Time Warping to discover patterns which are repeated in lifelog images. If this approach gives meaningful patterns, the results will be helpful information for lifelog researches (e.g. segmentation, clustering). We evaluated whether this approach extracts meaningful patterns. As a result, we found that it has potential for lifelog researches.

## Team Name

AKBL

## Subtasks

Lifelog Insights sub Task (LIT)

## Keywords

Lifelog, Spoken Term Discovery, Dynamic Time Warping

## 1. INTRODUCTION

Lifelog is logging personal life or experience by some sensors (e.g camera, microphone, GPS, etc.). Recently, device, used to logging, have miniaturized and improved functionality. Therefore, long time lifelog, such as logging all of life, have come to possible. However, it is difficult to use , because such lifelog consist of enormous data. Hence, many researchers have studied efficient way to use lifelog.

For example, Aizawa et al.[1] proposed lifelog video retrieval system using camera, GPS, brain wave, web history, etc. Ellis et al.[4] proposed creating an automatic diary from acoustic lifelog by segmentation and clustering based on acoustic features. Doherty et al.[3] showed lifelog images segmentation based on similarity between adjacent images. (similar to TextTiling[5])

In this paper, we attempt applying of technique, which have studied in field of speech processing, for lifelog images. In typical speech processing, speech are represented as time sequence of feature vector which extracted from speech in order of time. Whereas, lifelog images are time sequence of images. Therefore, lifelog images are represented as similar to speech, by transforming image into vector. Here, we consider that techniques, which have studied in field of speech processing, are available to lifelog images. In this paper, we consider an approach called STD, which discover word-like pattern from speech. STD finds word-like pattern as pattern which are repeated in multiple speech. Park et al.[8] proposed an approach based on Dynamic Time Warping (DTW). Jansen et al.[6] improved Park et al. approach using image processing. If patterns, which are repeated in multiple lifelog, are found by applying such approach, its results are probably useful for lifelog tasks. In this paper, we evaluate whether this approach extracts repeated patterns. First, we run STD system, then obtain the patterns as its results. Next, we observe directly those results and evaluate whether useful or not.

The organization of this paper is as follow. In section 2, we explain about STD in detail, then we show experimental results on section 3. In section 4 we describe future work. Finally, in section 5, we conclude.

## 2. SPOKEN TERM DISCOVERY

Spoken Term Discovery (STD) is an approach to discover word-like pattern from speech. In this section, we describe DTW based approach (called Segmental DTW[8]) used to our study.

### 2.1 Segmental DTW

This approach is variation of DTW. In typical DTW, for two time sequences $X = \{x_1, x_2, \cdots, x_M\}$ and $Y = \{y_1, y_2, \cdots, y_N\}$, optimal alignment between $X$ and $Y$ is calculated, and similarity of two sequences is measured. First, distance matrix $D \in \mathbb{R}^{M \times N}$ is created, each element $D(i, j)$ is corresponding to local distance $d(x_i, y_j)$ between $x_i$ and $y_j$. On this matrix $D$ between $D(1, 1)$ and $D(M, N)$, optimal alignment, minimizes cumulative distance, is found by dynamic programming. Obtained global alignment path can be used to measure similarity of two sequences. However,it is not suitable for finding word from utterance.

Park et al. proposed Segmental DTW for solve this problem. Segmental DTW consist of below procedure.

First, dividing distance matrix into diagonal overlapping bands with width $W$, and calculating alignment path within each band. These bands prevent undesirable alignment. Also, they allow for multiple alignments, as each band corresponds to another potential path with start and end points that differ from $D(1, 1)$ and $D(M, N)$.

Second, searching subsequence of alignment path, which have highest similarity, for each alignment path. Highest similarity means likely to be word. When searching subsequence, length of subsequence is constrained with minimum length $L$. The minimum length criterion is used to prevent spurious matches between short segments. Figure 1 shows

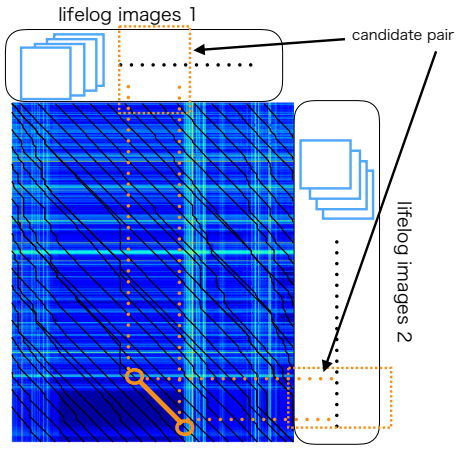


**Figure 1: Segmental DTW. This figure shows an example of multiple alignment on distance matrix. For each alignment path, we search best alignment (shown as orange line), then part of sequence (shown as orange rectangle), which associated with it alignment, are extracted as candidate for repeated pattern.**

an example this algorithm.

In original approach, clustering as post processing is performed. However, we didn't perform it, because our purpose is to evaluate whether this approach can extracts useful pattern. We will consider clustering in future work.

## 2.2 Application for Lifelog images

Here, we describe application Segmental DTW for lifelog.

### 2.2.1 *Vector representation for image*

As previously mentioned, we consider that lifelog images are represented as similar to speech, by transforming image into vector, and therefore speech processing approaches are available to lifelog images. Hence, we consider about vectorize for images firstly. In popular way, image features, such as SIFT, based Bag-of-Keypoints[2] is used. This is inspired by Bag-of-Words in NLP. However, we use visual concepts which are included by test collection. There are two reasons, first, creating Bag-of-Keypoints is costly than using visual concepts, second, we don't have a detailed knowledge of image processing.

### 2.2.2 *Distance*

DTW using distance for calculation, and therefore we must decide distance between vector. In this case, we use euclidean distance.

$$\|\boldsymbol{p}-\boldsymbol{q}\| = \sqrt{\sum_{i}^{n}(p_i - q_i)^2}, \tag{1}$$

where, $\boldsymbol{p} = \{p_1, p_2, \cdots, p_n\}$,$\boldsymbol{q} = \{q_1, q_2, \cdots, q_n\}$.

### 2.2.3 *Calculation of DTW*

DTW calculates cumulative distance on distance matrix to find optimal alignment. We use below formula.

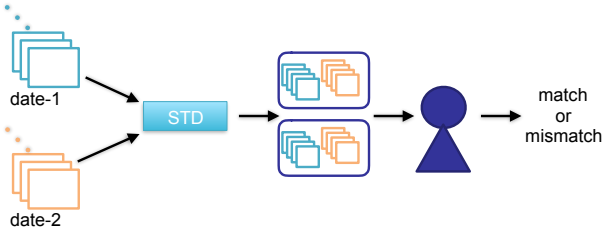


**Figure 2: Outline of experiment. First, we input two days lifelog images into STD system, then obtain candidate pairs. Next, we observe these pairs, and judge pairwise match or mismatch.**

$$W(i,j) = d(x_i, y_j) + min\left\{\begin{array}{l} W(i-1,j) \\ W(i,j-1) \\ W(i-1,j-1) \end{array}\right\}, \tag{2}$$

where, $d(x_i, y_j)$ is distance between $x_i$ and $y_j$ (in this case, euclidean distance), $W(i,j)$ is cumulative distance at $(i,j)$.

### 2.2.4 *Band width $W$ and minimum length $L$*

In speech processing, suitable value for these parameter is known empirically, but it is unknown in this case. We take not extreme value. About concrete value, we describe in Sec. 3.

In this way, we apply STD approach for lifelog images sequence. Given two lifelog images sequence as input, our system output candidates for patterns which are maybe shared by these input sequences. Each candidate is pair of subsequence of lifelog images. Then, we experiment to evaluate whether our approach can extracts shared patterns.

# 3. EXPERIMENT

In this section, we describe experiment which are performed to evaluate whether our approach can extracts patterns which are repeated in lifelog.

## 3.1 Experimental setup

We use a few days lifelog images from single user included in NTCIR-12 Lifelog task test collection[], because we considered that similar pattern is so few in case of using across user data. For parameters, we set window width $W$ to 100, and minimum length $L$ to 50. In our approach, window width corresponding to number of samples, and each sample is image. We investigate time interval between consecutive images on data used to our experiment. As a result, time interval is about 30 seconds, hence our window width means temporal width about 50 minutes. On the other hand, minimum length corresponding to length of alignment path. Note that alignment path length doesn't necessarily corresponding to time differ from window width.

As shown in Figure 1, candidates are obtained as pair of subsequence corresponding to best alignment on each alignment path. If our approach is useful for pattern extraction, the pair are expected to includes similar pattern. Here, we directly observe obtained pairs, then judge pairwise 'match' or 'mismatch', where 'match' means that sub-

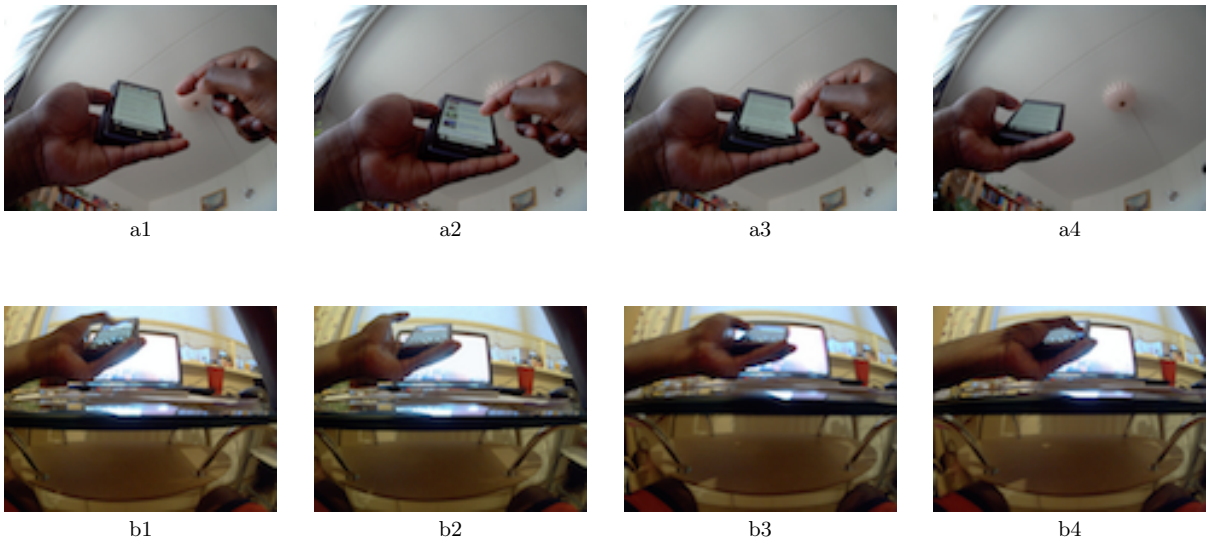


**Figure 3: Sample of 'match' pair.**

sequences, which are contained in candidate pair, to some extent includes similar pattern, and 'mismatch' means that subsequences no includes similar pattern. Figure 2 shows outline of our experiment.

In Figure 3,4, we show sample image pairs are extracted by STD. Fig. 3 is 'match' pair sample, and Fig. 4 is 'mismatch' pair sample. These images are a part of subsequence constructed by about 20 - 50 images. In Fig. 3, images shows similar action (in this case, "using smartphone"), while Fig. 4 images shows completely different scene (Fig. 4(a) shows 'watching TV', and Fig. 4(b) shows 'cleaning room').

We judge pair to be 'match', if both subsequences includes similar action, and judge to be 'mismatch' if not.

## 3.2 Results

**Table 1: Experimental results.**

| judge | pair |
|---|---|
| match | 103 |
| mismatch | 170 |

As shown in Table 1, our approach extracted 'mismatch' pair more than 'match' pair, unfortunately. However, 'mismatch' pair often has low DTW score, while 'match' pair has high DTW score. Therefore, we have considered that eliminating 'mismatch' pair from candidates by filtering with DTW score. By this filtering, our approach will probably works more effectively.

Reviewing lifelog images which were used to our experiment, we found 'match' patterns other than automatic extracted patterns. There are some reasons for that these patterns were not extracted. The one reason is minimum length parameter probably. The patterns which has length shorter than our parameter, were not extracted. This problem will be possibly improve, if we find optimal parameter. However, we have considered that too short pattern is unsuitable, similar to case of speech. (In case of speech, too short pattern means phoneme, not word.) Therefore, we also take into account that to ignore short pattern. Apart from short length, images of bad quality may also be reason. For example, images has blurred when user moves hard. Alternatively, the camera, which is used to lifelogging, is mounted on chest, so user's arm covers most of screen occasionally. The influence of bad quality may be strong, because we have used only image. By looking again 'match' pairs, we found that our system have extracted only patterns which are constructed by relatively small movement (e.g. 'using smartphone', 'using computer', 'watching TV', etc.). Also from this fact, we perceive influence of bad quality. (However, there is not much patterns that user moves hard, in images are used to experiment. So, it might be incorrect.) It might be difficult to avoid this problem, because we have used only images to extract patterns.

There are problems to be solved, but we believe that our approach has potential for lifelog tasks.

# 4. FUTURE WORK

In the future work, first, we will examine whether filtering based on DTW score is effectively for elimination of 'mismatch' pairs. If it is effectively, our approach works better. Next, we will optimize details of approach (e.g. vector representation, distance between vector, calculation of DTW alignment, etc.), because we have not examined yet. For example, we have considered that use of Bag-of-Keypoints instead of visual concepts, and introduction of slope constraint to alignment path of DTW.

Finally, we will consider useful application for lifelog. For example, we may be able to perform scene detection, by dividing lifelog images based on obtained patterns. Its segmentation enables to retrieve lifelog on small unit, such as event unit but not day unit. Also, we can do clustering for obtained patterns. In original Segmental DTW, Park et al. used graph clustering for word discovery. Lin et al.[7] reported Time Constrained K-means (TCK-means) clustering for lifelog video. TCK-means is extended K-means algo-

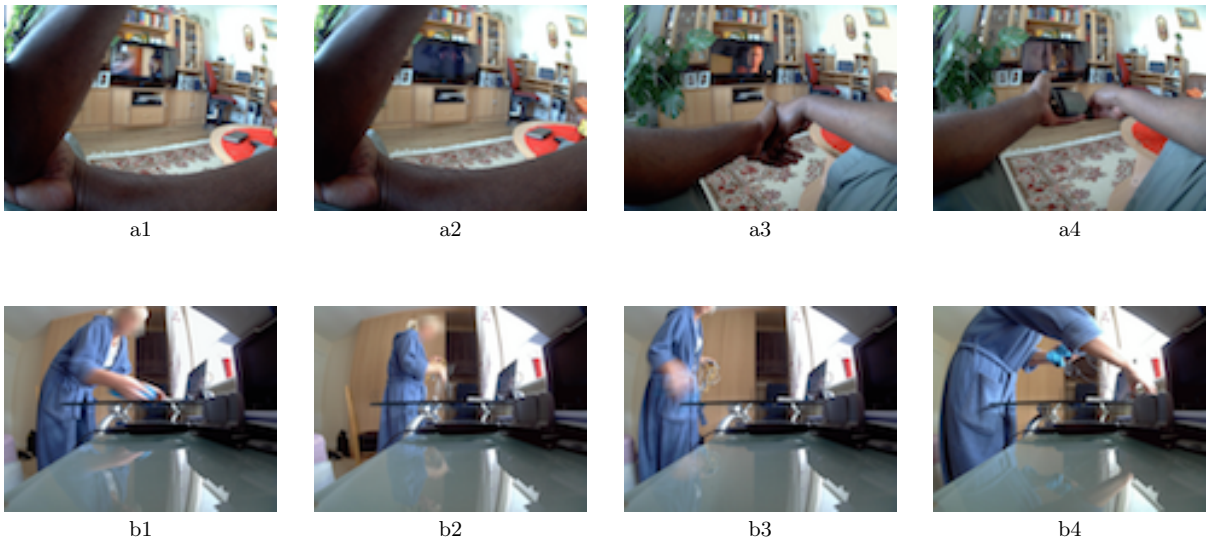


**Figure 4: Sample of 'mismatch' pair.**

rithm, it take into account temporal relationship between data. By clustering, patterns including similar action, are collected into group. We may obtain information from group, and use it for automatic labelling or automatic summary. And, if many patterns are associated by group, it may be available for retrieval (e.g. like pseudo relevance feedback).

As mentioned in Sec 1, these researches has been studied on various approach. We consider that comparison with preexist approaches and combination with those.

## 5. CONCLUSION

In this paper, we examined of application of Spoken Term Discovery technique,which have been studied on field of NLP, for lifelog images. We used unsupervised technique (called Segmental DTW) using variation of Dynamic Time Warping. We applied it by representing each lifelog image as vector. Then, we evaluated effectiveness of our approach, by to observe outputs of STD and to judge these outputs to be 'match' or 'mismatch'. As a result, we confirmed that our approach has potential for lifelog tasks, but we also found several problems.

In future work, to improve these problems, we plan to introduce filtering with DTW score and to optimize some part of approach. Then, we intend to examine various application.

## 6. REFERENCES

[1] K. Aizawa. Digitizing personal experiences: Capture and retrieval of life log. In *Multimedia Modelling Conference, 2005. MMM 2005. Proceedings of the 11th International*, pages 10–15. IEEE, 2005.

[2] G. Csurka, C. Dance, L. Fan, J. Willamowski, and C. Bray. Visual categorization with bags of keypoints. In *Workshop on statistical learning in computer vision, ECCV*, volume 1, pages 1–2. Prague, 2004.

[3] A. R. Doherty and A. F. Smeaton. Automatically segmenting lifelog data into events. In *Image Analysis for Multimedia Interactive Services, 2008. WIAMIS'08. Ninth International Workshop on*, pages 20–23. IEEE, 2008.

[4] D. P. Ellis and K. Lee. Minimal-impact audio-based personal archives. In *Proceedings of the the 1st ACM workshop on Continuous archival and retrieval of personal experiences*, pages 39–47. ACM, 2004.

[5] M. A. Hearst. Texttiling: Segmenting text into multi-paragraph subtopic passages. *Computational linguistics*, 23(1):33–64, 1997.

[6] A. Jansen, K. Church, and H. Hermansky. Towards spoken term discovery at scale with zero resources. In *INTERSPEECH*, pages 1676–1679, 2010.

[7] W.-H. Lin and A. Hauptmann. Structuring continuous video recordings of everyday life using time-constrained clustering. In *Electronic Imaging 2006*, pages 60730D–60730D. International Society for Optics and Photonics, 2006.

[8] A. S. Park and J. R. Glass. Unsupervised pattern discovery in speech. *Audio, Speech, and Language Processing, IEEE Transactions on*, 16(1):186–197, 2008.