

一种基于Lifelogging视频的文本标签生成模型

刘 洋, 刘国奇

沈阳建筑大学计算机科学与工程学院, 辽宁 沈阳

收稿日期: 2024年12月20日; 录用日期: 2025年1月16日; 发布日期: 2025年1月24日

摘 要

从2011年开始, 我们发起了一个收集个人信息生活记录数据的项目, 该项目收集了22位志愿者的4万条lifelogging数据。随着时间的推移, 志愿者的lifelogging数据越来越多, 其中收集到的视频就多达3020条, 想要搜索这些lifelogging数据中的视频变得非常困难。因此, 我们提出了一种视频分解 + 图像分析的模型Liu-VTM (Video Tags Model), 该模型从lifelogging视频中筛选能够代表该视频内容的关键帧, 并依据关键帧进行图像识别得到视频的标签, 最后可以通过标签直接检索到相应的视频。在本次实验中我们探讨了多种视频选取关键帧的方法对模型的影响, 并提出了一个新的评价指标“最佳内容覆盖率”用于评价lifelog领域内视频选取到的关键帧的性能。我们的实验结果证明了Liu-VTM模型可以有效对lifelogging数据集打上视频标签并依据标签直接检索到相应视频。

关键词

生活日志, 视频关键帧选取, 视频检索, 视频标签生成

A Text Label Generation Model Based on Lifelogging Videos

Yang Liu, Guoqi Liu

School of Computer Science and Engineering, Shenyang Jianzhu University, Shenyang Liaoning

Received: Dec. 20th, 2024; accepted: Jan. 16th, 2025; published: Jan. 24th, 2025

Abstract

Since 2011, we have initiated a project to collect personal lifelogging data, gathering 40,000 lifelogging entries from 22 volunteers. Over time, the amount of lifelogging data from the volunteers has increased, including as many as 3020 videos, making it extremely difficult to search through these lifelogging videos. Therefore, we propose a video decomposition and image analysis model called Liu-VTM (Video Tags Model). This model selects keyframes from lifelogging videos that represent the content of the video and uses image recognition on these keyframes to generate video tags.

These tags can then be used to directly retrieve the corresponding videos. In this experiment, we explored various methods for selecting keyframes from videos and proposed a new evaluation metric called “Optimal Content Coverage Rate” to assess the performance of keyframe selection in the lifelogging domain. Our experimental results demonstrate that the Liu-VTM model can effectively tag videos in lifelogging datasets and retrieve the corresponding videos based on these tags.

Keywords

Lifelogging, Video Keyframe Selection, Video Retrieval, Video Tagging

Copyright © 2025 by author(s) and Hans Publishers Inc.

This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

<http://creativecommons.org/licenses/by/4.0/>



Open Access

1. 引言

随着科学技术的发展, 个人数据渐渐走进大家的视野, 这类数据主要包括个人的健康数据、生活数据等, 我们称之为 Personal Big Data (PBD) [1] [2], Lifelogging 是 PBD 中的来源之一。Lifelogging 是指人们能够通过不同的细节程度对自己的日常生活进行数字化的记录[3]-[5], 并将其应用于各种场景, 实现各种目的[6]-[9]。Lifelogging 数据本身具有隐私的特性[10] [11], 并且 Lifelog 数据集中的数据是一直在进行增长的, 伴随着 lifelog 日志数量的增加, 如何快速浏览和管理这些数据是一个值得研究的问题[12]-[15]。

在过去的二十年里, 研究人员们提出了很多用于管理、挖掘和快速检索 Lifelogging 数据的技术。Gurrin 等人注重于 lifelog 领域内的数据管理, 并对 lifelongging 数据进行了数据存储的区分、组织和可视化[1]。Microsoft 通过将 lifelog 技术作用于医疗, 成功地帮助一名患有短期记忆丧失的人[16]。Harvey 等人[17]将心理学和 lifelog 相结合, 从心理学的角度分析了人类的记忆机制, 并提出了一种基于图形分割, 上下文增强(认识物体和人)和图像检索的增强记忆的方法。

Lifelog 领域内的图像信息绝大部分都是关于对照片的记录和分析。Byrne 等人首次进行了 lifelog 领域内物体识别的工作, 并验证了有监督的概念识别[18]。Venugopalan 等人和 Yao 等人[19] [20]都以第一人称时间叙述探索了图像的文本描述生成。但在 lifelog 领域中对于视频的分析和管理却少之又少, 一是因为早年间在 lifelog 领域内因为可穿戴设备和存储硬盘普遍偏大, 导致相较于存储一个短时间内的视频, 人们更倾向于存储一些尽可能覆盖全天的图片。这使得早期的 lifelog 数据集中视频数据普遍偏少。随着近年来计算机硬件性能的提升, 使得小型的可穿戴设备记录大量视频成为可能。

Lifelog 领域内的视频通常具有高频率、长时段记录[1]、个人视角[21]、隐私敏感[10] [11]、需要高效地存储和处理[22]、包含多种活动[23]的特点。Lifelogging 数据是一直在进行增长的, 随着视频数据的增加使得对视频进行检索变得十分困难。经典的视频检索技术是通过人工审阅的方式对视频中的信息进行识别并分配适当的标签, 再通过标签对视频进行检索[24]。因此, 分配的标签质量遵循主观标准并且因人而异。Khan 等人的初步实验[25]表明, 由于人类信息回忆能力缺乏精确性, 人类生成的元数据不足以全面洞察视频的主要内容。此外, 手动生成的语义标签过于缓慢并且存在不规则性。因此通过从视频中选取关键帧[26]-[28], 对关键帧进行语义标签的识别能够加速这一过程。基于此我们提出了 Liu-VTM 模型, 该模型采用视频分解 + 图像分析的结构, 分别对视频进行关键帧的选取、使用深度学习算法完成关键帧到标签的生成[29], 然后对标签进行清洗和筛选。最后我们将视频生成的标签添加到 lifelogging 数据中, 方便数据上传者能够快速检索到自己的视频。

本研究的总体结构分为六章, 包括引言一章。第二章首先介绍了我们 13 年间收集到的 Liu-Lifelog 数

据集, 并且列举了我们人工对这 3020 条视频日志进行详细的浏览并分类出 7 种不同主题的视频, 每个主题根据内容和视频规格选出了 2 或 3 个具有代表性的视频, 最后形成了一个拥有 16 个视频 Liu-lifelog 标准视频集。第三章介绍了作者提出的 Liu-VTM 模型和“最佳内容比”的概念。第四章使用我们的 Liu-Lifelog 数据集中的视频对 Liu-VTM 模型继续验证并对实验结果进行详细的分析。第五章建立视频检索系统, 用于展示通过文字搜索来查找视频日志。最后一章对全文进行总结, 并展望未来的研究方向。

2. 相关工作

2.1. Liu Lifelog Project

Liu Lifelog project started from 2011 是我们团队的一个公开的 APP, 现在任何人都可以在我们的网站上免费下载这个应用, 从而参加到我们的项目中。任何人都可以免费地获取我们已经公开的 Lifelog 数据。在过去 13 年中, 我们拥有超过 20 个志愿者参与到 Liu Lifelog Project 之中, 其中志愿者 Liu 的数据集多达 12000 多条日志, 每条日志都包含图片、文字、地理位置信息。

在 LifeLog 移动端上, 用户注册登录后展示的页面如图 1 所示, 用户可以拍摄一张照片或者视频, 并选择此刻在进行着什么行为活动, 最后将完整的信息上传。图像或视频形式的记录不但能够记录当前活动的状态, 还可以记录当下的心情。当用户想要搜索日志信息的时候, 只能通过文字搜索来查询到其中某个时刻的文字记录。

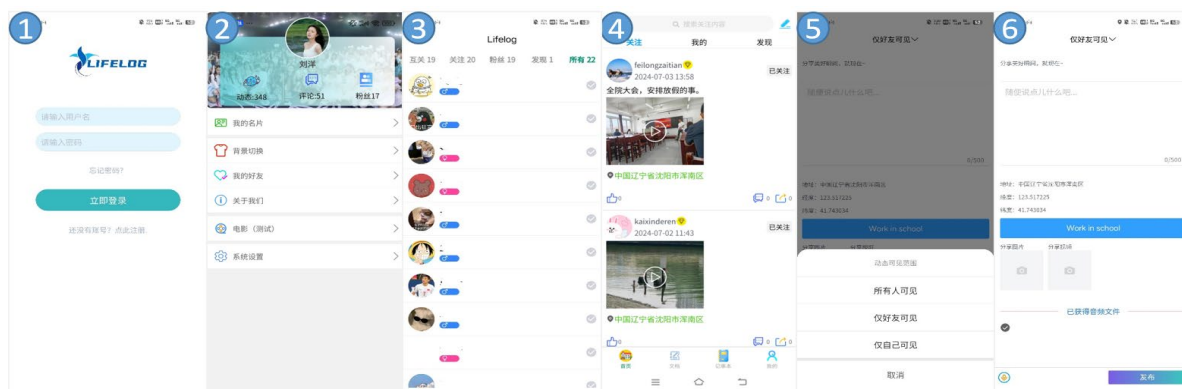


Figure 1. System display page
图 1. 系统展示页面

2.2. Liu-Lifelog 数据集

Liu-Lifelog 数据集(<http://www.lifelog.vip/>)是 Liu 团队花费 13 年时间记录的数据。该数据集有以下优势: 采样时间长, 参与者人数众多, 数据分布于全国各个城市且持续更新。每天有十几名用户分享他们的日常生活, 使这个数据集成为生活日志数据的综合存储库。其中一条日志记录可以通过时间、经纬度、行为、地点、文本描述、图片、视频一共七种数据类型构成。其中包含视频日志记录多达 3020 条, 部分数据展示如图 2 所示。

我们通过对所有包含视频日志进行浏览并结合 Liu-Lifelog Project 的行为分类发现 Liu-Lifelog 数据集中的视频存在不同的主题, 作者通过对 3020 条视频数据进行逐条整理并分类出了 7 种不同主题的视频, 每个主题根据内容和视频规格选出了 2 个或 3 个具有代表性的视频, 形成了一个拥有 16 个视频的 Liu-lifelog 标准视频集, 其中 On Road 和 Recreation 因为其种类多样从而选择了 3 个具有代表性的视频, 标准视频集数据如图 3 所示。

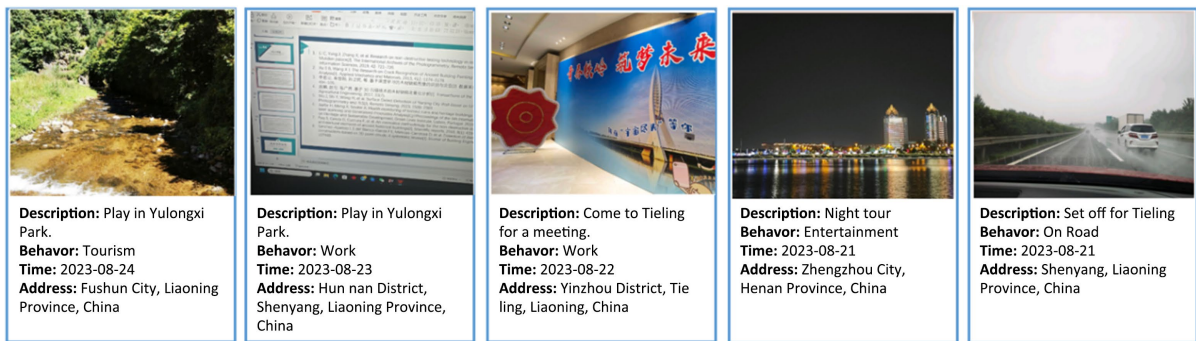


Figure 2. Example of lifelog data

图 2. 日志数据样例

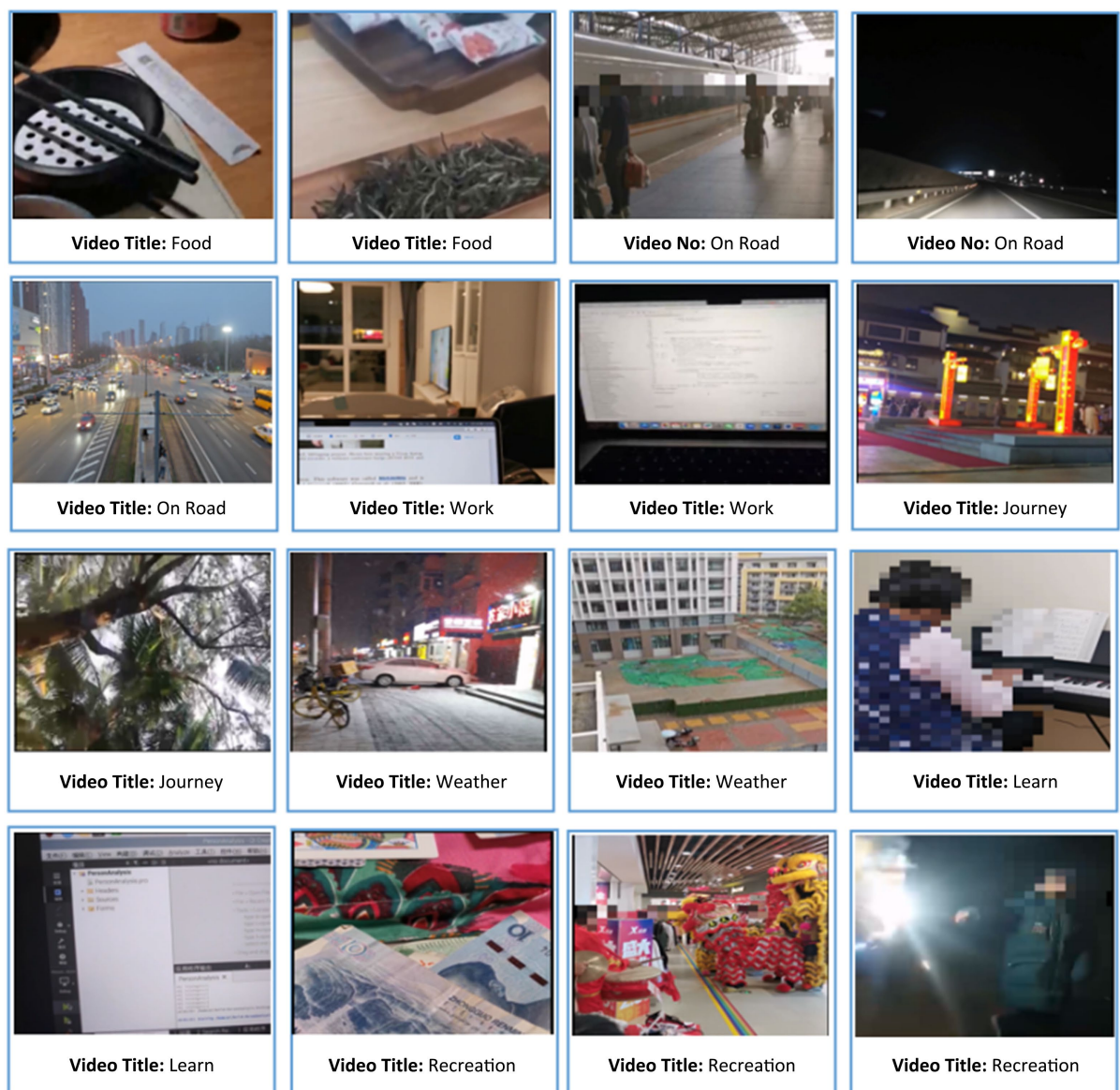


Figure 3. Display of the standard dataset

图 3. 标准数据集展示

3. Liu-VTM 模型

3.1. Liu-VTM 模型阐述

为了解决 Lifelog 领域中视频生成标签的问题, 我们提出了一个 Liu-VTM 模型, 该模型通过将视频处理成图片, 并对图片进行标签生成来实现视频生成标签, Liu-VTM 模型分为四个部分, 分别是视频数据处理、视频帧选取、帧内容标签生成以及标签分析及筛选, 模型如图 4 所示:

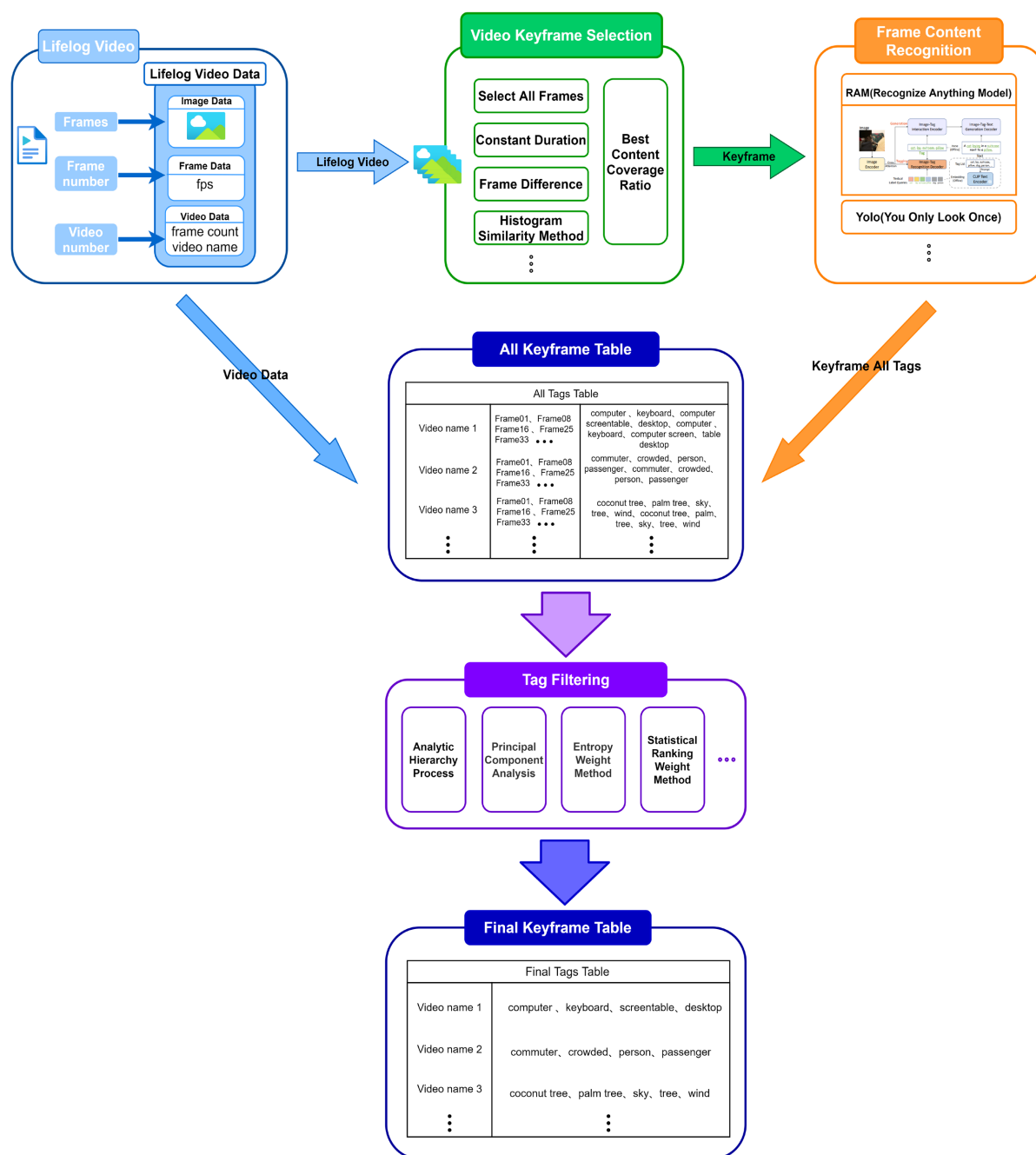


Figure 4. Liu-VTM model
图 4. Liu-VTM 模型图

视频数据处理是为了得到视频中包含的图像帧、视频帧率、视频总帧数这三个数据。其中视频中的多个图像帧的图像信息可以反映出该视频的内容主题, 视频帧率和总帧率可以用于反映视频的帧数量。

视频帧选取是为了构造适用于该视频集的最佳内容比, 并在最佳内容比的选取下找到最适用于该视频集的视频关键帧选取算法。对于一个视频数据集而言, 不同的视频关键帧的选取算法所选出的结果也不同, 这也就代表不同的关键帧算法对于该视频数据集的视频选取效果有好坏之分, 为了能够评价不同的算法对视频关键帧选取的好坏, 作者在模型中加入了最佳内容比的概念。通过最佳内容比我们可以比较不同选帧算法对于同一个视频集中视频关键帧选取的优劣, 从而可以得到一个适用于该视频集的视频选帧的算法。我们将最佳内容比与视频关键帧算法一起作为 Liu-VTM 模型中视频帧选取的模块。

帧内容标签生成是对视频关键帧进行内容识别, 封装视频内容识别标签。通过视频帧选取后, 将选取出的视频关键帧输入到帧内容标签生成模块中, 通过图像识别算法如 RAM、YOLO 等主流的图像内容识别算法, 将识别出的图像内容、帧位置、视频编号等封装成一条标签数据, 由多条标签数据最后汇聚成一个巨大的视频标签表。

标签分析及筛选是对视频内容标签进行清洗, 并赋权。将这个视频标签表输入到标签筛选模块进行该视频集的标签赋权和标签清洗, 在帧内容标签生成模块中, 生成的图像标签中会存在很多相同的内容标签、以及识别错误的内容标签, 针对于这种情况, 我们对视频标签表中相同视频编号下的内容标签继续标签赋权, 通过视频标签权重的不同, 可以将权重过低的图像标签认定为错误的内容标签, 将其进行去除。并且我们也可以通过使用标签权值的大小, 来获取该视频的主要描述内容。

3.2. 最佳内容覆盖率的观念

视频选取关键帧的过程是从视频中选取一组能够代表视频内容的静态图像。Lifelogging 视频数据通常记录了拍摄者的一段生活经历, 这一段生活经历可能包含多段事件, 这不利于直接使用视频进行分析, 而是要先将其分割成多个代表不同事件的切片然后对视频进行编号。这样对于单个视频中的选取多个关键帧本质上就是代表了尽可能对这个独立事件进行更加详细的概括。限于每个人视频记录的长短问题, 使用固定的关键帧数量去选择视频中的关键帧, 所带来的弊端就是最后筛选出的关键帧容易发生对于短的视频而言重复性很高, 对于长的视频而言并不能够概括长视频的内容这种情况。因此, 我们选择了根据视频内容而选择不同数量的关键帧。

针对视频如何选择根据画面选择不同数量的关键帧, 我们提出了一个“最佳内容覆盖率”的评价指标值, 用于评定最后选择的关键帧数量尽可能少的情况下反映视频的主要内容, 所使用的用于评估视频帧内容覆盖率的公式如下:

$$R = \frac{K}{N} \quad (1)$$

R 为压缩率, 定义为关键帧数量与视频帧数量的比例, 其值也在 $[0, 1]$ 范围内, N 为原视频中的总帧数, K 为最终所选定的关键帧数量。

$$S = w_C \cdot C + w_R \cdot (1 - R) \quad (2)$$

S 为最佳内容覆盖率, C 为覆盖率, 其值的范围为 $[0, 1]$, 其中 1 表示完全覆盖, w_C 和 w_R 分别为覆盖率和压缩率的权重, 并且 $w_C + w_R = 1$ 。

其中 $1 - R$ 表示为考虑压缩效果的同时反映信息保留的程度, 因为较低的压缩率意味着保留了更多的原始信息。当我们更加注重覆盖率时, 那么其中 $w_C > w_R$; 反之如果, 是为了尽可能的减少关键帧的数量

而可以牺牲一些覆盖率, 那么可以选择 $w_C < w_R$ 。最佳内容覆盖率可以非常清晰的展示我们使用不同的方法找出的不同数量的关键帧对于视频内容的好坏程度。

针对不同的视频数据集会有不同的最佳内容覆盖率, 因为 lifelog 领域中的视频对于关键帧的选择是高度主观的, 所以需要使用先使用主观选择法, 从同一个视频数据集中选出部分能够代表该视频集的视频, 并对这些视频进行主观选帧。然后对主观所选的视频帧进行压缩率和内容覆盖率的计算, 并将这两个值带入到上述公式(2)中。使用熵权法计算出独属于该视频的最佳内容覆盖率, 信息熵计算公式如下:

$$e_j = \frac{1}{\ln(m)} \sum_{i=1}^m p_{ij} \ln(p_{ij}) \quad (3)$$

m 是样本数, p_{ij} 是第 i 个样本在第 j 个指标上进行标准化后的数据, e_j 是第 j 个指标的信息熵。然后使用信息熵进行每个指标的权重 w_j 的计算, 权重计算公式如下:

$$w_j = \frac{1 - e_j}{\sum_{j=1}^n 1 - e_j} \quad (4)$$

n 是指标的总数, $\sum_{j=1}^n 1 - e_j$ 是所有指标的信息效用值的总和。

通过熵权法我们可以计算出对应视频集的权重 w_C 和 w_R , 将其带入到公式(2)中就可以得到独属于该视频集的最佳内容覆盖率计算公式, 然后使用不同的视频选帧算法计算主观选帧时使用的视频, 从而得到不同算法的最佳内容覆盖率, 当不同视频选帧的最佳内容覆盖率与主观选帧的最佳内容覆盖率越接近, 则说明该视频选帧算法对于该视频集越好。

4. 实验过程

4.1. 数据预处理

Liu Lifelog Project 在早期因为硬件服务器条件的限制以及早期系统构造人员的技术水平局限, 导致在 Liu-Lifelog 数据集中的视频出现了编码以及存储问题, 其表现为存储视频出现一个视频只有一帧的画面、原本的一个视频因为压缩和传输的问题被转为多个只有两三帧的视频、视频被过度压缩出现的视频画面分辨率大幅度下降已经到了看不清无法进行分析的地步等情况, 作者对全部的 3020 条视频中的这些视频分别进行处理, 将重复的视频数据进行删除只保留下一条, 将不能进行分析的视频以及只有一两帧的视频进行清理, 最后得到能够使用的视频还有 2703 条(截止到 2023 年 8 月)。

4.2. 适用于 Liu-LifeLog 的最佳内容覆盖率

在 Lifelog 领域中的视频数据通常记录了拍摄者的一段生活日程, 得益于 Liu-Lifelog Project 的日志型 Lifelog 收集方式, 我们团队收集的一个日志中的视频都只代表了一个单独的事件, 对于进行数据筛选后的 2703 个视频数据而言, 这 2703 个视频数据代表了 2703 个独立事件。那么对于单个视频中选取多个关键帧, 本质上就是代表了尽可能对这个独立事件进行详细的概括。我们的视频数据因为每个人对日志的记录习惯的问题, 导致其中有的视频只有几秒钟, 有的视频却多达十几分钟甚至更多。所以我们最后选择了根据视频内容来选择不同数量的关键帧。

视频中对于关键帧的选择是高度主观的, 由于我们数据集中的视频拆解成图像后的数量过于庞大, 所以笔者设计了一个实验框架用于确定 Liu-Lifelog 里最佳内容覆盖率中的权值。笔者根据 Liu-Lifelog 标准视频数据集中的 16 个视频, 从 2703 个视频中再选出 32 个具有代表性的视频数据, 将这 48 个视频都拆解为图像, 最后得到 28,110 张图像, 并让发布者从其中挑选出能够代表对应视频的关键帧。为了防止发布者对于大量图像筛选出现懈怠和晕厥等反应, 笔者让发布者在选择一个视频中的关键帧后间隔几天

进行恢复, 并且发布者在选择视频关键帧之前并不知道自己要选择的视频内容。为了进行综合评定, 在发布者选择后, 笔者对所选关键帧进行高斯滤波器和拉普拉斯算子相结合的方式来确定最后发布者选择的关键帧质量是否合格。

通过以上的方​​式将主观选择出的关键帧进行覆盖率和压缩率的计算并使用熵权法计算出权重 w_C 和 w_R , 将权重 w_C 和 w_R 带入到公式 2 中求得符合 Liu-Lifelog 视频数据的最佳内容覆盖率公式。

4.3. 视频关键帧选取方法比较

我们使用了四种方法进行视频帧的选取, 第一种是逐帧选择法, 第二种是直方图相似度比对选取相似度较小的视频帧, 第三种是使用了间隔取帧法, 第四种则为帧间法来获取视频帧, 以下为四种视频帧选取方法的选帧策略。

1) 直方图相似度对比法: 直方图相似度对比法是将该视频中的视频帧依次进行直方图相似度对比选取其中相似度低的视频帧作为描述视频的关键帧。

2) 间隔取帧法: 使用固定的间隔时间将该视频中的视频帧进行取出, 本文中采用了间隔 24 帧/30 帧/60 帧/120 帧, 通过读取视频的帧率来决定间隔一秒钟所取得图像, 最终将隔秒取出的视频帧作为描述视频的关键帧。

3) 帧间法: 帧间差分法通过依次计算每两帧之间的帧间差分, 进而得到平均帧间差分强度, 最后选择具有平均帧间差分强度局部最大值的视频帧作为描述视频的关键帧。

4) 逐帧选择法: 逐帧选择法是将该视频中的每一个视频帧都取出作为描述视频的关键帧。

我们基于选出的关键帧计算不同视频对应方法下的最佳内容覆盖率值, 并与主观选择视频帧的方法计算出的最佳内容覆盖率值进行对比。下表 1 只详细列出了 16 个标准视频不同选取方法对应的最佳内容覆盖率的分数。

Table 1. The best content coverage rate scores for standard videos

表 1. 标准视频最佳内容覆盖率分数

视频序号	主观选择帧	直方图相似法	间隔取帧法	帧间法	逐帧法取帧
1	71.47	62.48	94.41	61.66	36.70
2	74.65	70.05	95.33	76.50	36.70
3	64.63	62.59	95.84	70.16	36.70
4	91.81	88.46	96.04	91.49	36.70
5	67.37	58.93	79.47	60.03	36.70
6	79.35	75.38	74.34	77.42	36.70
7	82.15	76.17	77.08	61.04	36.70
8	68.22	68.70	95.82	62.11	36.70
9	63.61	63.79	69.63	62.87	35.95
10	63.37	62.63	96.70	64.18	35.55
11	68.67	62.56	94.71	64.68	36.70
12	80.00	96.99	96.99	98.59	36.70
13	72.79	77.70	97.68	66.55	36.70
14	73.72	66.20	96.79	70.58	36.70
15	65.75	62.91	71.05	62.12	35.03
16	67.09	70.42	94.98	67.26	36.70

我们使用以上方法对 2703 个视频进行视频取帧, 最终得到逐帧法图像 2,232,700 张, 直方图相似度对比法图像 10,990 张, 间隔取帧法图像 46,050 张, 帧间法图像 28,250 张。基于选出的视频帧计算不同视频对应方法下的最佳内容覆盖率并进行详细的分析。

表 1 中的分数都是根据 Liu-Lifelog 视频集中根据主观选择法所计算出的最佳覆盖率得到的, 其中的分数是根据与主观选择法的方差大小作为评判指标, 也即是同一视频的情况下与主观选择法的分数差距越小越好, 如图 5 所示为其余方法与主观选择法之间的方差图。

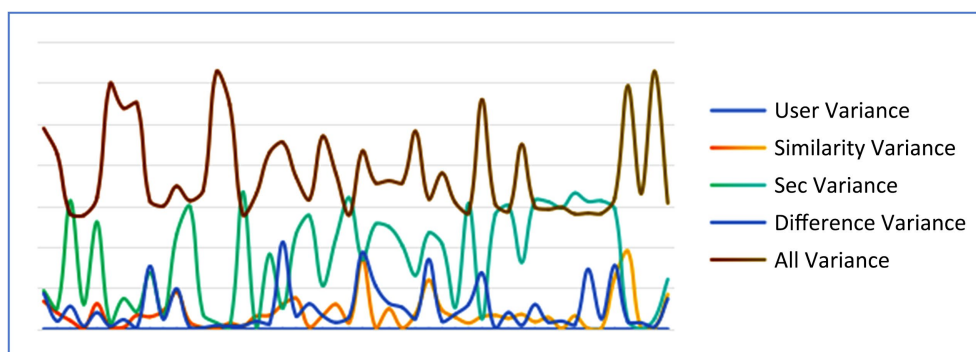


Figure 5. Comparison variance graph with subjective selection
图 5. 与主观选择法比较的方差图

因为逐帧法是每一帧都取出作为关键帧所以逐帧法的压缩率为 1, 从而得到逐帧法的关键帧选择效果是 4 种方法中最差的一种, 从图 5 中可以看到逐帧法是距离主观选择法最远的。间隔取帧法相较于逐帧选择法而言, 其压缩率更高(小于 1, 压缩率越低越好), 但是因为间隔取帧法并没有对内容进行分析, 所以其在内容覆盖上会存在很多相似的部分, 方差图上表现得要比逐帧法好但是比帧间差分法和直方图相似法都要差。

帧间差分法的效果与直方图相似法的效果是最接近的, 两者对于视频关键帧的选取非常接近主观选择法的选择结果, 但是相较于直方图相似法而言, 帧间差分法是通过求得相邻两帧之间图像对应位置的像素值差的绝对值, 判断其是否大于某一阈值从而判断两帧之间的相似度, 相较于直方图相似法在计算中会对遍历过的图像中相似度较低的视频帧进行保留而言, 帧间差分法对应视频内容的判断要差一些, 其内容覆盖率相较于直方图相似法而言会更低, 所以得到直方图相似法是最适用于 Liu-Lifelog 视频日志的视频关键帧选择方法。所以笔者在系统搭建中使用直方图相似法来对视频进行关键帧的选择。针对于不同的视频日志体系需要进行具体的评定从而确定更加适合自己的视频选帧方法。

4.4. 实验结果

通过对不同的视频取帧方法进行最佳内容覆盖率的计算和比较最终选定直方图相似法作为本系统的视频关键帧选取方法。在确定选取方法后, 我们借鉴了 oppo 实验室使用的 RAM (Recognize Anything Model) 进行视频关键帧的内容识别。RAM 可以识别生活中常见的物体类别, 识别准确度高, 并且其引入了一种新的图像标记的范例, 即通过利用大规模的图像文本用以进行训练而不是手动注释。相较于 yolo, RAM 在物品识别方面展示出了非常强大的性能。

结合 RAM, 我们将 Liu-Lifelog 中的视频导入到模型中, 通过使用多种不同的视频选帧方法的使用并通过最佳内容覆盖率的筛选, 最终确定一个视频选帧的方法并将选出的视频帧作为描述视频的关键帧输入到 RAM 中, 得到对于该视频的内容标签表。

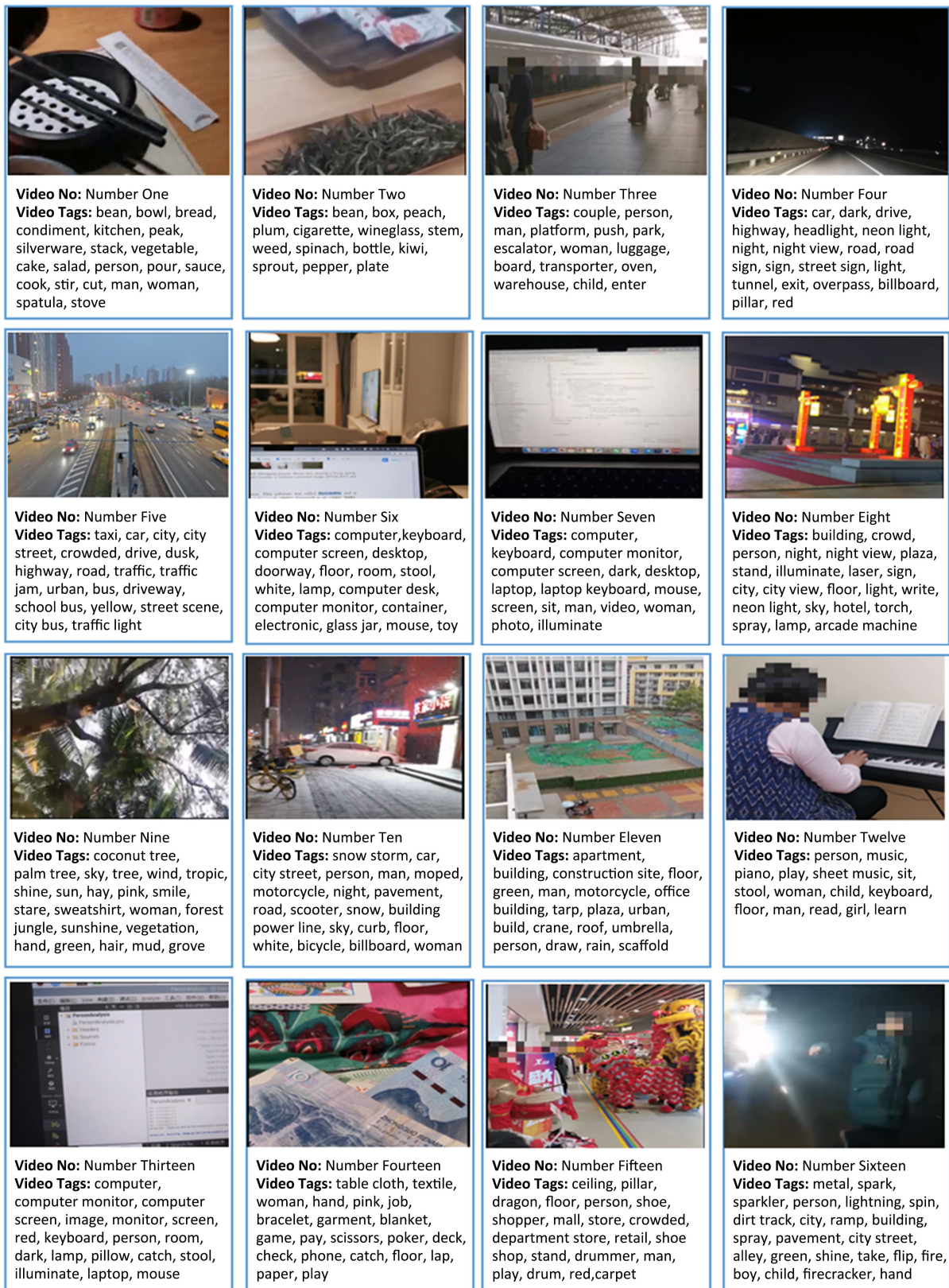


Figure 6. Partial display of standard video tags
图 6. 标准视频标签部分展示

基于该内容标签表, 我们根据出现次数进行赋权, 并将其中只出现一两次的标签进行清洗。最后得到对于 16 个视频的标签如图 6, 其中只展示主要的标签部分, 其余视频的标签展现在我们的网站中。

5. 系统实现

5.1. 系统框架

Lifelog 一个重要的功能就是回忆, 通过查看 lifelogging 数据, 用户能想起许多过去珍贵的记忆。我们基于 web 的应用程序检索 lifelogging, 用户可以根据需求检索想要回忆的内容。在本节中, 我们根据需求将 Liu-VTM 模型加入到 Liu-lifelog 的需求检索系统中去, 通过 Liu-VTM 模型用户可以按照自己的需求搜索到视频内容。系统的流程图如图 7 所示。

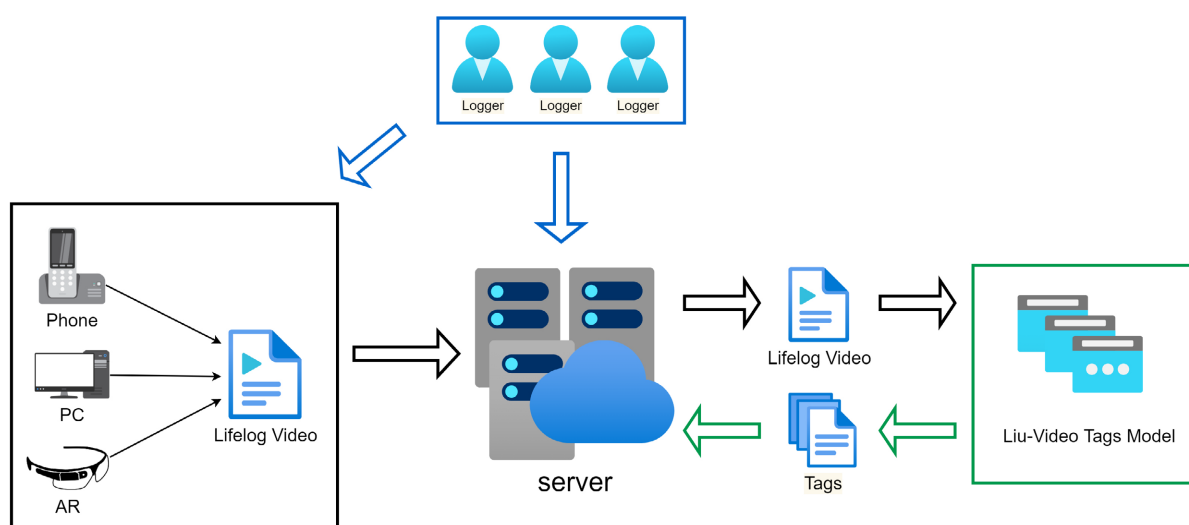


Figure 7. System flowchart with Liu-VTM integrated

图 7. 加入了 Liu-VTM 的系统流程图

Logger 通过手机、PC 和 VR 等硬件设备记录 Lifelog Video, 并将其上传到 Liu-Lifelog project 的服务器中, Liu-VTM 模型会对上传的 Lifelog Video 进行定时的标签生成, 并将其添加到 Lifelogging 数据中。此时, Logger 可以通过搜索自己对于之前发过日志内容中的物体或者行为, 来查找想要回忆的内容。这种根据 Logger 经历过的事件日志中内容来进行检索的方式更加符合人类的记忆联想方式。通过该系统, 我们可以清晰直观地查看 Logger 过去的一些事情。对于数据上传者而言, 这是非常有意义的, 因为该系统可以帮助他们想起自己过去的记忆。

5.2. 系统展示

我们使用 Liu-VTM 模型添加到 Liu-lifelog 系统中的展示页面, 如图 8 所示。我们将 Liu-VTM 模型所生成的中文标签和英文标签都展示到了 tags 和 tags (english) 中去。其中生成的标签是 Liu-VTM 基于左侧 Photo 中的视频进行生成和筛选出来的。

我们可以看到针对同一条 lifelog 数据会存在多个不同的视频事件, 其相对应的中文标签和英文标签也会相应地增加。我们可以通过左上角的搜索功能, 通过回忆视频中的物件或者行为, 能够快速地搜索到与之对应的 lifelog 数据。这样是更加接近人类本身的记忆检索的方式。

- [8] Nguyen, T., Le, T., Ninh, V., Tran, M., Thanh Binh, N., Healy, G., *et al.* (2021) Lifeseeker 3.0: An Interactive Lifelog Search Engine for Lsc'21. *Proceedings of the 4th Annual on Lifelog Search Challenge*, Taipei, 21 August 2021, 41-46. <https://doi.org/10.1145/3463948.3469065>
- [9] Tran, L., Kennedy, D., Zhou, L., Nguyen, B. and Gurrin, C. (2022) A Virtual Reality Reminiscence Interface for Personal Lifelogs. In: Jónsson, B., *et al.*, Eds., *MultiMedia Modeling*, Springer International Publishing, 479-484. https://doi.org/10.1007/978-3-030-98355-0_42
- [10] Ksibi, A., Alluhaidan, A.S.D., Salhi, A. and El-Rahman, S.A. (2021) Overview of Lifelogging: Current Challenges and Advances. *IEEE Access*, **9**, 62630-62641. <https://doi.org/10.1109/access.2021.3073469>
- [11] Liu, G., Rehman, M.U. and Wu, Y. (2021) Personal Trajectory Analysis Based on Informative Lifelogging. *Multimedia Tools and Applications*, **80**, 22177-22191. <https://doi.org/10.1007/s11042-021-10755-w>
- [12] Khan, I., Ali, S. and Khusro, S. (2019) Smartphone-Based Lifelogging: An Investigation of Data Volume Generation Strength of Smartphone Sensors. In: Song, H.B. and Jiang, D.D., Eds., *Simulation Tools and Techniques*, Springer International Publishing, 63-73. https://doi.org/10.1007/978-3-030-32216-8_6
- [13] Ribeiro, R., Neves, A. and Oliveira, J.L. (2020) Image Selection Based on Low Level Properties for Lifelog Moment Retrieval. *12th International Conference on Machine Vision (ICMV 2019)*, Amsterdam, 16-18 November 2019, 9-18. <https://doi.org/10.1117/12.2557073>
- [14] Xu, Q., Molino, A.G.D., Lin, J., Fang, F., Subbaraju, V., Li, L., *et al.* (2021) Lifelog Image Retrieval Based on Semantic Relevance Mapping. *ACM Transactions on Multimedia Computing, Communications, and Applications*, **17**, 1-18. <https://doi.org/10.1145/3446209>
- [15] Ali, S., Khusro, S., Khan, A. and Khan, H. (2021) Smartphone-Based Lifelogging: Toward Realization of Personal Big Data. In: Guarda, T., *et al.*, Eds., *Information and Knowledge in Internet of Things*, Springer International Publishing, 249-309. https://doi.org/10.1007/978-3-030-75123-4_12
- [16] Hodges, S., Williams, L., Berry, E., Izadi, S., Srinivasan, J., Butler, A., *et al.* (2006) Sensecam: A Retrospective Memory Aid. *8th International Conference, UbiComp 2006*, Orange County, 17-21 September 2006, 177-193. https://doi.org/10.1007/11853565_11
- [17] Harvey, M., Langheinrich, M. and Ward, G. (2016) Remembering through Lifelogging: A Survey of Human Memory Augmentation. *Pervasive and Mobile Computing*, **27**, 14-26. <https://doi.org/10.1016/j.pmcj.2015.12.002>
- [18] Byrne, D., Doherty, A.R., Snoek, C.G.M., Jones, G.J.F. and Smeaton, A.F. (2009) Everyday Concept Detection in Visual Lifelogs: Validation, Relationships and Trends. *Multimedia Tools and Applications*, **49**, 119-144. <https://doi.org/10.1007/s11042-009-0403-8>
- [19] Venugopalan, S., Rohrbach, M., Donahue, J., Mooney, R., Darrell, T. and Saenko, K. (2015) Sequence to Sequence—Video to Text. *2015 IEEE International Conference on Computer Vision (ICCV)*, Santiago, 7-13 December 2015, 4534-4542. <https://doi.org/10.1109/iccv.2015.515>
- [20] Yao, L., Torabi, A., Cho, K., Ballas, N., Pal, C., Larochelle, H., *et al.* (2015) Describing Videos by Exploiting Temporal Structure. *2015 IEEE International Conference on Computer Vision (ICCV)*, Santiago, 7-13 December 2015, 4507-4515. <https://doi.org/10.1109/iccv.2015.512>
- [21] Doherty, A.R. and Smeaton, A.F. (2008) Automatically Segmenting Lifelog Data into Events. *2008 9th International Workshop on Image Analysis for Multimedia Interactive Services*, Klagenfurt, 7-9 May 2008, 20-23. <https://doi.org/10.1109/wiamis.2008.32>
- [22] Gemmell, J., Bell, G. and Lueder, R. (2006) MyLifeBits: A Personal Database for Everything. *Communications of the ACM*, **49**, 88-95. <https://doi.org/10.1145/1107458.1107460>
- [23] Aizawa, K., Hori, T., Kawasaki, S. and Ishikawa, T. (2004) Capture and Efficient Retrieval of Life Log. *Pervasive 2004 Workshop on Memory and Sharing Experiences*, Vienna, 20 April 2004, 15-20.
- [24] Zhou, L., Hinbarji, Z., Dang-Nguyen, D. and Gurrin, C. (2018) Lifer: An Interactive Lifelog Retrieval System. *Proceedings of the 2018 ACM Workshop on The Lifelog Search Challenge*, Yokohama, 11 June 2018, 9-14. <https://doi.org/10.1145/3210539.3210542>
- [25] Khan, U.A., Ejaz, N., Martinez-del-Amor, M.A. and Sparenberg, H. (2017) Movies Tags Extraction Using Deep Learning. *2017 14th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, Lecce, 29 August-1 September 2017, 1-6. <https://doi.org/10.1109/avss.2017.8078459>
- [26] Gibson, D., Campbell, N. and Thomas, B. (2002) Visual Abstraction of Wildlife Footage Using Gaussian Mixture Models and the Minimum Description Length Criterion. *2002 International Conference on Pattern Recognition*, Vol. 2, 814-817. <https://doi.org/10.1109/icpr.2002.1048427>
- [27] Truong, B.T. and Venkatesh, S. (2007) Video Abstraction: A Systematic Review and Classification. *ACM Transactions on Multimedia Computing, Communications, and Applications*, **3**, 3-es. <https://doi.org/10.1145/1198302.1198305>

- [28] Lv, C. and Huang, Y. (2018) Effective Keyframe Extraction from Personal Video by Using Nearest Neighbor Clustering. 2018 *11th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI)*, Beijing, 13-15 October 2018, 1-4. <https://doi.org/10.1109/cisp-bmei.2018.8633207>
- [29] Ilyas, S. and Ur Rehman, H. (2019) A Deep Learning Based Approach for Precise Video Tagging. 2019 *15th International Conference on Emerging Technologies (ICET)*, Peshawar, 2-3 December 2019, 1-6. <https://doi.org/10.1109/icet48972.2019.8994567>