Proceedings of the 15th International Conference on Mobile and Ubiquitous Multimedia (MUM '16), December 12–15, 2016, Rovaniemi, Finland

Remind - Towards a Personal Remembrance Search Engine for Motion Augmented Multi-Media Recordings

Philipp M. Scholl
Albert-Ludwigs-University
Freiburg, Germany
pscholl@ese.uni-freiburg.de

Kristof van Laerhoven
Albert-Ludwigs-University
Freiburg, Germany
pscholl@ese.uni-freiburg.de

ABSTRACT

A searchable database of multi-media recordings provides a way to augment one's memory. This database might contain video, audio and motion data, which is indexed to allow for quick searches. Ultimately, queries for similarity on each recorded modality would be supported. For example video sequencing showing similar objects or comparable sequences of gestures can be retrieved. An important aspect of this challenge is how to encode such multi-modal data, and how to make it searchable. One approach, based on a multi-media container format, is proposed in this paper together with an architecture to allow for similarity queries on multiple modalities.

Author Keywords

multi-media; motion; indexing; encoding; querying

ACM Classification Keywords

H.2.4 Informations Systems: Database ManagementSystems-Multimedia Databases; H.5.m. Information Interfaces and Presentation (e.g. HCI): Miscellaneous

INTRODUCTION

Remembering details about a manual task, even shortly after executing it, can be challenging - especially when it is not possible to document at hand. Wearable systems, like Google Glass, smart watches and cameras, could potentially alleviate this situation. By recording video, audio and other modalities, the user can be supported in documenting and looking up information during the task at hand. This can involve executing experiments in wet labs, cooking in a kitchen, assembling products or maintaining machinery. One cue into such recordings could be the sequence of motions or gestures executed during such a task. While these can be queried for explicitly by keywords, one could also imagine a system in which the last few minutes of recorded motion is used to query a database of similar motion. This paper is concerned with a way to encode such data to facilitate such queries.

MUM '16 December 12-15, 2016, Rovaniemi, Finland

ACM ISBN 978-1-4503-4860-7/16/12.

DOI: http://dx.doi.org/10.1145/3012709.3016074

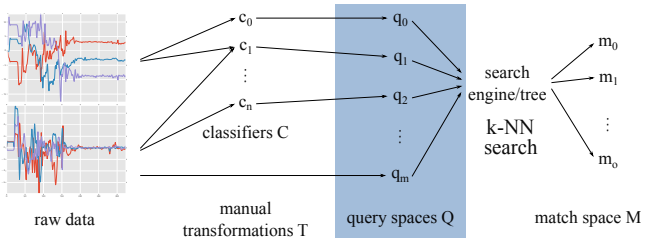

Figure 1. The actual similarity of raw motion data, which is used for indexing the multi-media files is fully defined by a mapping into a query space Q. This can either be done manually, or with the help of machine learning algorithms. The mapped sequence of data can then again be encoded into the multi-media file, and indexed by a search tree for fast k-Nearest Neighbour queries.

Each sensor generates a stream of data, in a certain format, at a certain rate and with parameters that have to be stored in order to compare them to other recording sessions. All of these recording parameters should be stored in conjunction with the actual sensor data. Additionally, recording multiple sensors also involves synchronization. Even when running on the same device, independent sensors operate on their own clock, which renders time synchronization an important issue. For example, an accelerometer will deliver its samples faster, than a light sensor, if its package is placed in the vicinity of a heat source. Therefore, even system-local sensors, need to be synchronized on a common clock, in the same way multiple recording devices on a network need to synchronized on a common clock.

Despite those issues, organizing recordings of such an ensemble of wearable sensor quickly becomes a burden. One possible approach is to use a multi-media container to store synchronized sensor streams on a common time-axis, together with their recording parameters. The matroska standard defines such a file format [5]. In this format, multiple video streams can be stored, as well as multiple audio, subtitle and data streams. These streams can be compressed with state-of-the-art codecs. Motion data can be stored as an audio stream as well, storing some of its recording parameters like sampling rate and format. Optionally it can be compressed with a lossless audio codec. This provides a standard way to encode recording parameters, as well as the data itself in a synchronized way. Recording sessions, or rather memories, of a user can then be stored as such a multi-media file.

The idea of providing a searchable database of personal recordings has been researched in the area life-logging already. In-

Sensed [1], MyLifeBits [4], and Stuff I've Seen [3] are recent and early examples of systems, which provide a database to help in remembering details about one's personal life. A survey of recent tools and approaches to this challenge are given in [2]. From a psychological point of view, one could also ask which cue might benefit the user best, which is partly answered [6]. In contrast to these works, this paper addresses the technical challenge of storing "memories". Which might involve multiple media streams, as well as indexing them in an open format which can be exchanged and used by different applications.

The question however is how to render these multi-media files searchable. How can we provide a way to query these for similar gestures, both by name and by the raw data recorded during different sessions? For example, how can we enable a biologist to execute a "pipetting" gesture and find recordings where she was executing a similar gesture. Both by querying the systems manually, in its literal sense by moving her arm, and by querying explicitly by a keyword like "pipetting". In the remainder of this paper, we will describe a possible approach to this problem.

ENCODING THE SIMILARITY OF MOTION

The problem of querying for similar motion and by descriptive keywords can be viewed as finding a mapping from the space of raw sensor data R to a query space Q, in which a clear definition of similarity is encoded (see Fig. 1). Activity Recognition is one research field which is concerned with how such a mapping can be defined and found. Approaches range from manually defining this mapping to fully automated machine learning systems, where only a limited set of parameters are pre-selected. The input for such a system is always labelled raw data, and for unknown raw data the output is an element of Q (or label if you like). The input labels, as well as the output labels are bound to regions in the continuous sensor streams. These can be encoded into the matroska document as subtitles, which encodes a start and end timestamp, as well as a corresponding label, exactly the output of aforementioned mapping.

Put into the context of the biologist's example, the query space Q consists of labels such as "pipetting", "mixing" and other manual actions. Once a mapping from the raw data R, which map sequences of such data to elements of Q, is found, it can be used to generate subtitles for the recording. This means a gesture recognition system, which maps movement data from R to labels such as "pipetting" of Q, can be encoded as a subtitle. These subtitles are then stored side-by-side with the multi-media recording, and used as a possible search cue. The encoding of such subtitles is general enough to span sequences of fixed and varying sizes, and can also be used for hierachical label sets.

FAST SEARCH ON METRIC SPACES

Under which conditions is a database of such classified recordings quickly searchable? This depends on the nature of the query space Q, and the similarity measure g defined on it. If g is metric, a *metric tree* data structure can be used to accelerate the search on Q. If it does not fulfill this criteria, only a linear search can be used. Complexity-wise, a tree-accelerated search query can be executed in $\mathcal{O}(\log n)$, while a linear search has a complexity of $\mathcal{O}(n)$, where n is the database size.

It is important to note, that similarity of patterns might not be fully defined by above mentioned classifiers/mappings. This is only the case, if the similarity measure on Q is an exact match. Since metric tree algorithms support fast Nearest Neighbour (NN) lookups, a k-NN type of classification can be built on top of Q again. Which in case of inexact matches, allows for defining another tier of classification, i.e. the definition of similarity can be recursed. The result, however, is always a start and end time when a similar motion was executed.

CONCLUSION

One important design decision for a wearable remembrance agent is the format in which sensor data is stored. Here, we proposed an approach based on a multi-media container, which encompasses all relevant parameters of a recording session. Furthermore, assuming that a number of classifiers for motion data is available, the classification results are to be stored side-by-side with the sensor data. If no such classifier is available, a similarity mapping would have to be defined manually (e.g. euclidean distance), or enough labelled data would have to be collected to build such a classifier. The container format renders the exchange of data easier and by having a standard format, allows for easier development of novel classification systems. The results of these classifications can then also be quickly searched, and by proper choice of similarity measure also accelerated. With such a system in place, the user might not only easily document his manual task, but also quickly query for similar recording by his recent motions.

REFERENCES

1. M. Blum, A. Pentland, G. Troester, and Z.S. ETH. 2005. InSensed: Wearable Collection of Multimedia based on Interest. *Master's Thesis* 599 (2005).
2. Tilman Dingler, Passant El Agroudy, Huy Viet Le, Albrecht Schmidt, Evangelos Niforatos, Agon Bexheti, and Marc Langheinrich. 2016. Multimedia memory cues for augmenting human memory. *IEEE Multimed.* 23, 2 (2016), 4–11.
3. Susan Dumais, Edward Cutrell, J J Cadiz, Gavin Jancke, Raman Sarin, and Daniel C Robbins. 2003. Stuff I've seen: a system for personal information retrieval and re-use. *Proc. 26th Annu. Int. ACM SIGIR Conf. Res. Dev. informaion Retr.* 49, 2 (2003), 72–79.
4. Jim Gemmell, Gordon Bell, Roger Lueder, Steven Drucker, and Curtis Wong. 2002. MyLifeBits: fulfilling the Memex vision. *Proc. tenth ACM Int. Conf. Multimed.* (2002), 235–238.
5. Non-Profit Organization Matroska. 2016. The Matroska File Format. (2016). `https://www.matroska.org/`
6. Elise van den Hoven and Berry Eggen. 2014. The cue is key: Design for real-life remembering. *Zeitschrift fur Psychol. / J. Psychol.* 222, 2 (2014), 110–117.