Automated Multimedia Diaries of Mobile Device Users Need Summarization

Marc Gelgon, Kevin Tilhou

Automated multimedia diaries of mobile device users need summarization

M. Gelgon and K. Tilhou

IRIN / Ecole polytechnique de l'université de Nantes
rue C.Pauc, La Chantrerie, 44306 Nantes cedex 3, France
Tel : (33) 2 40 68 30 14 Fax : (33) 2 40 68 30 66
marc.gelgon@polytech.univ-nantes.fr

Abstract. This paper addresses a still original issue and a solution that, while emerging from the pattern recognition point of view, certainly shares common goals with mobile HCI research goals. The contribution is at the crossroads of multimedia data analysis for content-based retrieval, and wearable computing. As users are acquiring multimedia content personal mobile devices, they are getting also undergoing information overflow. The problem of structuring the content into time-oriented meaningful episodes is addressed, and we argue that geographical location processing is crucial, as a complement to processing audiovisual material. A technique for model-based temporal structuring of one's trajectory during a day is presented, based on a Bayesian/MAP approach, that generates one or several summaries. Experimental results illustrate the applicative interest of the problem addressed and validates the proposed solution

1 Context and objective

The work exposed in the present paper addresses needs related to information retrieval in personal mobile devices, especially aiming at those which purpose is to be continuously carried by their user, such as mobile phones and PDAs. Indeed, such small wireless terminals are undergoing considerable progress in their ability to capture, store, process and transmit data, notably of the multimedia form. Several models that incorporate image and audio acquisition, as well as geolocation of the users and that are targeting the general public, are currently being released, or are soon to be on the market.

The viewpoint of the paper is from the pattern recognition community rather than human-computer communication, but the goal of the work certainly coincides with HCI preoccupations and its purpose is cross-fertilization of ideas. As multimedia content is gathered, it progressively builds up a valuable memory of ones's life, which can be later searched, whether for practical, emotional, or much-sought fun pass-time purposes. A means of access that could prevail is the familiar time management software available on PDAs, with calendar-like

views on which multimedia content could be "annotated". Yet, for the owner to quickly, reliably and comfortably retrieve a well-defined piece of information in a large collection, or merely browse in it to get an overall idea, the content ought to be organized, at least partly automatically. More precisely, bearing in mind the the stringent input and display constraints of mobile devices, we believe a crux is the ability to generate compact summaries of one's activities during a given period, that suit visualization and browsing needs, including the above-mentioned need with the straying-type of browsing. Our focus in on proposing techniques for structuring the collection along the time-axis, via analysis of the multimedia data.

Although the more general problem of content-based retrieval of audiovisual material has largely being addressed for several years (e.g. TV summaries[20, 9]), understanding of needs and working out of solutions dedicated to the the particular context tackled here remains rather open. Early exploration of the automated diary concept was proposed in [13, 18], yet leaving out multimedia. Recently, proposals have been made towards organizing one's personal image collection, based on image features [14]. Towards organization of video collected from a wearable computer, a summarization technique was presented in [2], which rates the interest of video shots via measurements on brain waves. The structures that are extracted from the personal data collection (e.g. through statistical learning) are utilized here for navigating within the collection, but they may also benefit other purposes : context-awareness [1] and prediction [6]. For instance in [15], the authors propose automatic discovery of the user's frequented places and assignement of location-sensitive reminders. Alternatively, a technique to learn and infer location from image sequences acquired from a worn camera is presented in [21], so as to avoid the difficulties of accurate indoor localization. Interestingly, a paper is being published in parallel to the present [3], that addresses very similar goals (automatic multiscale determination of one's pertinent locations), although aiming at prediction rather than browsing the past.

In [8], we proposed to determine, from wearable video, periods when the user had met people, by coupling probabilistic face detection with HMM, leading to visual time-oriented summaries. In the present paper, we examine the unsupervised generation of summaries from geographical position sequence recordings. The results section shows how these contributions may be combined. We advocate the use of location because it appears realistic and useful. Indeed, it is not restricted to wearable computing platforms, but is also to be available on less intrusive, lower-price, ordinary mobile phones, whether with audio-visual capabilities of not. Further, it is relatively reliable, compared to most information that may be extracted from wearable images and audio. With time and identity of people the user met, it is likely to be the major search/browse criterion for searching one's data collection, and it is indeed quite easy to express for the user (e.g. compared to image-based queries). Also, from our experiments, location accuracy fades more slowly in one's memory than alternative search/browse cri-

teria, especially in the long run. Finally, the use of positioning only places modest storage and processing requirements.

Positioning technologies, surveyed in [7], may be GPS (possibly GSM network-assisted GPS), GSM basestation triangulation-based techniques such as E-OTD or its WCDMA successor, or cell identifier ; there is likely to be time-switching or fusion among these sensors, depending on local availability, reliability or subscription. Besides, indoor performance is under encouraging R&D [10] and post-processing is likely to be performed (e.g. in the form of enhanced Kalman filtering). Due to this improving and unstable situation, the present work is not dedicated to a particular technology or noise characteristics, but remains on the white Gaussian noise assumption (whereas actual corruption on measurements is likely to be time-correlated and its variance would be time-varying). Still, the statistical framework employed is flexible enough to introduce more elaborate noise models. Let us finally mention that in our application, the sampling rate would typically be a few seconds.

In the remainder of the paper, we formalize the problem and outline its solution (section 2), providing developments in sections (2.1) and (2.2). Section 3 presents experimental results, while section 4 summarizes the contribution and suggests further work.

2 Statistical trajectory segmentation

The general problem may be stated as summarizing the observed spatio-temporal trajectory $o = \{o_t = (x_t, y_t)\}, t = 1, .., T$ of the user. In fact, the sub-goal tackled in this paper consists in partitioning the time-series into time segments, represented by the sequence $s = \{s_t\}, t = 1, .., T$ of hidden state labels, in which s_t indicates to which of the M models o_t is assigned. The process should be as data-driven (unsupervised) as possible, regarding M the number of segments. With [16], and in contrast to [12], we consider that the degree of homogeneity within a segment should also be as much as possible estimated from the data, rather than be set a priori. Besides, we conjecture that segmentation of the trajectory according to a piece-wise parametric model provides a reasonnably meaningful account of the episodes that compose the period to be processed, at least for short periods (e.g. a day). Let $\Theta = \{\Theta_k, k \in \{1, \dots M\}\}$ denote the set of parameters vectors. Linear models are assumed in the remainder of the paper. We consider batch processing (that would occur e.g. every evening), but the technique proposed can be made incremental. Given these requirements, the problem comprises the two following aspects :

- the classical interwoven issues of unsupervised data clustering are gathered : estimating the model parameters, associating the data to the models, and determining the adequate number of models.
- the sequentiality of measurements should be introduced, so as to guarantee time-connexity of data assigned to a model, and limit spurious models due to very noisy measurements.

Denoting by S be the set of possible partitions, the chosen optimality criterion is the maximum a posteriori (MAP) label configuration, defined by (1).

$$\widehat{s} = \arg\max_{s \in S} p_\Theta(s|o) = \arg\max_{s \in S} p_\Theta(o|s)p(s) \quad (1)$$

The MAP criterion is a Bayesian estimator that transforms the search into an optimization problem. In our case, the search is conducted as follows : the search space is partitionned into subspaces $\{S_k\}, 1 \leq k \leq T$, where S_k contains all segmentations composed of k segments. Search within each subspace is conducted independently, as detailled in section 2.1. The solutions obtained for these various segmentation complexities are compared (section 2.2) and one or several global solutions are finally selected. For our application, we might not be only interested in a single, but in extracting several pertinent segmentations of the trajectory, especially if they can all summarize well the data, with various degrees of granularity, so as to supply the user with several alternative visual representations at multiple temporal scales, i accordion summarization [5].

2.1 Search for an optimal segmentation within a subspace S_k

The search for a maximum likelihood estimate of Θ is conducted by means of the Expectation-Maximization (EM) algorithm [17]. It is a well-known iterative scheme which, provided with some initialization for Θ, replaces the (ignored) data-to-model assignments by their statistical expectation *(E step)*, given the current parameter values. Therefrom, new parameter values may be computed *(M-step)*, until (garanteed, but possibly slow) convergence. In our case, still, the data-to-model assignments are more restricted as in common clustering situations, as data assigned to a model should be connected in time. In terms of our model, the probabilities of assigning a data element to a model are not independent among the data set. This structural constraint on the search space is dealt with by setting a hidden Markov model structure on the label sequence that constraints, via the transition matrix, connectivity of identical labels. Conditionally to models parameters Θ and segmentation complexity k, the MAP sequence is straightforwardly estimated with the Viterbi algorithm (hereunder 'MAP step'), which replaces our E step in the EM algorithm. As these models parameters Θ are unknown, iterations between MAP and M steps, which is known to convergence fast in practice, could be a solution but, because of the "hard" data-to-model assignements it carries out (in constrast with EM), it is known to lead to poorer parameter estimates than EM [4], which we confirmed experimentally. To combine EM and MAP-M so that their respective shortcomings are addressed, we employ the following three-phase scheme *(algorithm 1)*, in which results of each phase initialize the next one.

2.2 Model comparison and selection

The attempt to fit (e.g. linear) models to the data is to be viewed as the identification of the major trends and changepoints in the trajectory, that serve the

Algorithm 1

$\widehat{s^k_{init}} \leftarrow$ `segmentation in` k `segments of equal length`
$\Rightarrow$ a reasonnable initial guess
Do `MAP-M phase (until convergence)`

- `M-step : computer a maximum likelihood estimate of model parameters` Θ `(linear regression)`
- `MAP-step : estimate a MAP segmentation`

Loop
$\Rightarrow$ fast computation of rough parameter estimates, that supply the EM phase just below with a good initialization
Do `EM phase (10 iterations)`

- `M-step : estimate a maximum likelihood of model parameters` Θ `(weighted linear regression)`
- `E-step : compute the expectation of data-to-model assignements`

Loop
$\Rightarrow$ provides a better initialisation for the the MAP-M phase, thus leading to a more reliable segmentation
Do `MAP-M phase (until convergence)`

- `M-step : maximum likelihood estimation of model parameters` Θ `(linear regression)`
- `MAP-step : estimate a MAP segmentation`

Loop
$\Rightarrow$ final segmentation $\widehat{s^k}$ with k segments and likelihood $p(o|\widehat{s^k})$.

purpose of building a summary. A difficulty is that the more complex (flexible) the overall model proposed, i.e. the more segments we allow, the better it can fit to the data (e.g. in terms of likelihood), eventually undesirably fitting small local deviations due to noise. There exist principled approaches to defining the trade-off between goodness-to-fit vs. segmentation complexity, a problem well-known as Ockham's razor. The statistical view chosen is a well-founded framework for handling this unsupervised segmentation issue. A closely related, alternative, framework to handle this same issue is that of the information-theoretic viewpoint [11, 19]. Definition (1) in fact naturally exhibits this property. Let S^k be the set of all segmentations composed of k segments. All segmentation within S^k being*a priori* equal , the prior probability distribution $p(s^k)$ is flat and $p(s^k) = 1/card(S^k)$. This can be interpreted as the general Bayesian self-penalizing property of model complexity. Stirling's approximation conveniently re-expresses this result as $ln\, p(s^k) \simeq Tln(t) - (T-k)ln(T-k)$.

The search for the optimal segmentation according to (1) can thus be carried out as defined in *Algorithm 2*. The outcome of algorithm 1 is a set of candi-

Algorithm 2

`For all possible segmentation complexities` $k \in [k_{min}, k_{max}]$
Do
`search for the optimal segmentation` $\widehat{s^k}$ `using algorithm 1 , this provides` $ln\, p(o|\widehat{s^k})$.
`compute the complexity penalty` $ln\, p(\widehat{s^k})$
`compute the` *a posteriori* `probability` $ln\, p(\widehat{s^k}|o) = ln\, p(o|\widehat{s^k}) + ln\, p(\widehat{s^k}) + const.$
Loop

date segmentations (one per segmentation complexity) with their *a posteriori* probability.

An advantage retained from the Bayesian inference viewpoint is that, once provided with a posterior distribution, one is free to take decisions that suits one's dedicated needs, including extracting several pertinent solutions. The technique proposed for extracting *a set of segmentations* is to retain, among the solutions found in each of the subspaces, those which are local maxima, along the k-axis, within this set of solution. That is, if the trajectory inherently exhibits multiple levels of granularity (e.g. one moves between cities, but also within these cities), these levels can automatically be determined. Besides, the number of pertinent segmentations that can well explain the data is also automatically determined by the data, and is at least one. Let us finally underline that parametrization of the technique is low.

3 Experimental results

We report experimental results from GPS measurements, on which white Gaussian noise is added to simulate alternative sensors. The period reported corre-

sponds to activities during one day (going from home to two stores within a shopping center, then to work, temporarily to some university closeby, then to a library and finally back home. Fig 1 shows the corresponding trajectory. The two segmentations found to be relevant by the proposed scheme are overlaid (with some offset for readability), and both correspond to desirable outputs. The (minus) likelihood (penalized by segmentation complexity) for the optimal solution in each subspace S_k are plotted vs. k. Two local minima are identified. k_{max} is limited to 15, as finest summaries would anyway be unsuitable for the small GUI targeted. Fig. 2 shows a calendar-view displaying in parallel summaries from the present technique (left) and the ones based on face-detection (right). Images help summarize the face-based segments [8]. Interestingly, besides provding a good overview of the day, as the user has partial memory of the day, she can often identify the location via the face of the person that was met there, or vice-versa.

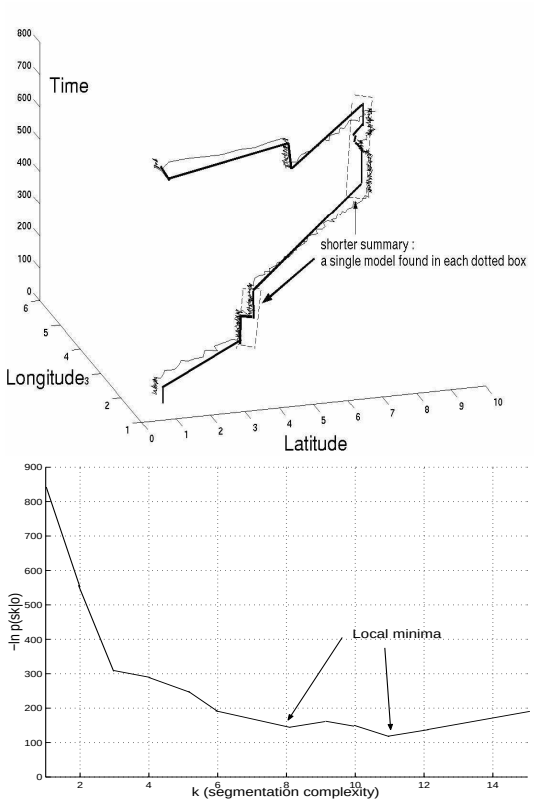

Fig. 1. Above: : spatio-temporal trajectory of the user, with the two-level summaries determined (the finest scale in bold lines; the coarsest resembles the finest, safe two grouping represented by dashed boxes).**Below:** penalized likelihood of candidate optimal segmentation with various complexities, and indication of the selected segmentations.

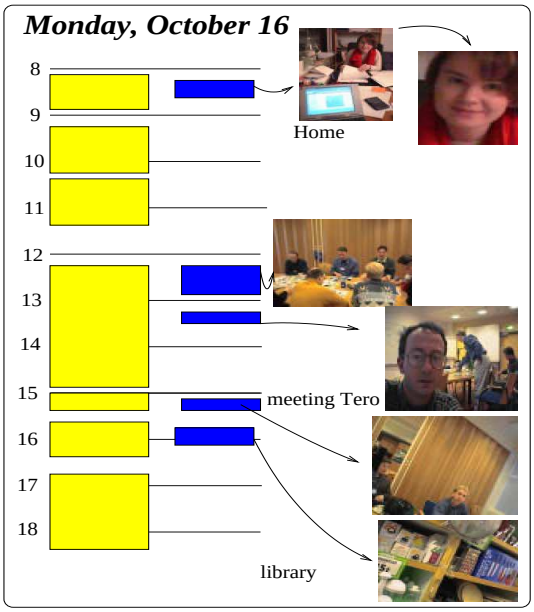

Fig. 2. Calendar-type of view constructed from the summaries built (the fine-scale location-based summary is chosen, left, and "meeting"-based summary (using face detection), right). Vertically is time of the day. Upon selection of a meeting, the face of the person met constitute an iconic summary. Images can similarly help indicate the location.

4 Conclusion

This paper has presented a contribution in a still rather open applicative field : personal multimedia data analysis for automatic annotation of electronic diaries. Arguing in favour of the use of unsupervised structuring of one's trajectory towards compact summaries to be displayed on a PDA screen, a statistical technique is proposed that, thanks to its Bayesian/MAP principle, provides an automatically determined number of time-oriented segmentations. This formalism would also easily enable GUI-experts, by rating the desirable complexity of summaries, to define and integrate a more elaborate prior probability distribution on the segmentation complexity. Appropriate naming of segments could partly come from a Geographical Information System, yet opening catchy issues. Finally, for longer durations, while the system is well able to separate "in town" and "out of town" remote trips, it has limitations in its "compression ability", due to the current summarization criterion. We are currently examining complementary criteria, based on detection and representation of periodicity, of usual vs. unusual.

Acknowledgements The authors are grateful to J.Jomppanen, A.Myka and J.Yrjänäinen from Nokia Research Center, Nokia corp., Finland, for discussions related to this work.

References

1. G.D. Abowd and E.D. Mynatt. Charting past, present and future research in ubiquitous computing. *ACM Trans. on Computer-Human Interaction*, 7(1):29–58, March 2000.
2. K. Aizawa, K. Ishijima, and M. Shiina. Summarizing wearable video. In *IEEE. Int. Conf. on Image Processing (ICIP'2001)*, pages 453–457, Thessaloniki, Greece, september 2001.
3. D. Ashbrook and T. Starner. Learning significant locations and predicting user movement with gps. In *To appear in IEEE Int. Symp. on Wearable Computer*, oct 2002.
4. C. Biernacki, G. Celeux, and G. Govaert. Strategies for getting the largest likelihood in mixture models. In *ASA Joint Statistical Meeting JSM 2000, invited paper*, Indianapolis, USA, 2000.
5. O Buyukkokten, H. Garcia-Molina, and A. Paepcke. Accordion summarization for end-game browsing on pdas and cellular phones. In *Proc. of ACM Computer Human Interaction (CHI'2001)*, Seattle, Washington, USA, 2001.
6. B. Clarkson and A. Pentland. Predicting daily behavior via wearable sensors. Technical Report Vismod TR 451, MIT, July 2001.
7. G.M. Djuknic and R.E. Richton. Geolocation and assisted GPS. *IEEE Computer*, pages 123–125, February 2001.
8. M. Gelgon. Using face detection for browsing personal slow video in a small terminal and worn camera context. In *IEEE. Int. Conf. on Image Processing (ICIP'2001)*, pages 1062–1065, Thessaloniki, Greece, september 2001.
9. M. Gelgon and P. Bouthemy. Determining a structured spatio-temporal representation of video content for efficient visualisation and indexing. In *5th European Conference on Computer Vision (ECCV'98), LNCS 1406-1407*, pages 595–609 (II), Freiburg, Germany, June 1998.
10. T. Haddrell and T. Pratt. Understanding the indoor GPS signal. In *Institute Of Navigation ION - GPS 2001 Conference*, pages 123–129, Salt Lake City, USA, September 2001.
11. M.H. Hansen and B. Yu. Model selection and the principle of minimum description length. *Journal of the American Statistical Association (JASA)*, 96(454):746–774, 2001.
12. E. Keogh, S. Chu, D. Hart, and M. Pazzani. An online algorithm for segmenting time-series. In *IEEE International Conference on Data Mining*, Silicon Valley, USA, December 2001.
13. M. Lamming and M. Flynn. Forget-me-not : intimate computing in support of human memory. In *Procs. of Friend21: Int. Symp. on next generation human interface*, pages 124–128, Meguro Gajoen, Japan, 1994.
14. J. Luo, A.E. Savakis, and A. Etz, S. Singhal. On the application of Bayes networks to semantic understanding of consumer photographs. In *IEEE. Int. Conf. on Image Processing (ICIP'2000)*, pages 802–807, Vancouver, Canada, september 2000.
15. N. Marmasse and C. Schmandt. Location-aware information delivery. In *IEEE Int. Symposium on handheld and ubiquitous computing*, pages 157–171, Bristol, U.K., sep 2000.

16. J. Oliver and C. Forbes. Bayesian approaches to segmenting a simple time series. In *Proc. of Proceedings of the Econometric Society Australasian Meeting, C. L. Skeels (ed), Canberry*, 1998.
17. R.A. Redner and H.F Walker. Mixture densities, maximum likelihood and the EM algorithm. *Society for Industrial and Applied Mathematics - SIAM Review*, 26(2):195–239, 1984.
18. B. Rhodes. The wearable rememberance agent : a system for augmented memory. *Personal Technologies Journal - Special issue on wearable computing*, 1(4):218–224, 1997.
19. S.J. Roberts, R. Everson, and I. Rezek. Minimum entropy data partitioning. *Proc. International Conference on Artificial Neural Networks*, 2:844–849, 1999.
20. M.A Smith and T. Kanade. Video skimming and characterization through the combination of image and language understanding techniques. In *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition*, pages 775–781, Puerto-Rico, juin 1997.
21. T. Starner, B. Schiele, and A. Pentland. Visual contextual awareness in wearable computing. In *IEEE Int. Symp. on Wearable Computing*, 1998.