Improving the Classification of Quantified Self Activities and Behaviour Using a Fisher Kernel

Peng Wang
National Laboratory for Information Science and Technology
Department of Computer Science and Technology
Tsinghua University
Haidian District, Beijing, China
pwang@tsinghua.edu.cn

Shiqiang Yang
National Laboratory for Information Science and Technology
Department of Computer Science and Technology
Tsinghua University
Haidian District, Beijing, China
yangshq@tsinghua.edu.cn

Lifeng Sun
National Laboratory for Information Science and Technology
Department of Computer Science and Technology
Tsinghua University
Haidian District, Beijing, China
sunlf@tsinghua.edu.cn

Alan F. Smeaton
Insight Centre for Data Analytics
Dublin City University
Glasnevin, Dublin 9, Ireland
alan.smeaton@dcu.ie

Ubicomp/ISWC'15 Adjunct, September 07–11, 2015, Osaka, Japan
 ISBN 978-1-4503-3575-1/15/09...$15.00
DOI: http://dx.doi.org/10.1145/2800835.2800947

Abstract
Visual recording of everyday human activities and behaviour over the long term is now feasible and with the widespread use of wearable devices embedded with cameras this offers the potential to gain real insights into wearers' activities and behaviour. To date we have concentrated on automatically detecting semantic concepts from within visual lifelogs yet identifying human activities from such lifelogged images or videos is still a major challenge if we are to use lifelogs to maximum benefit. In this paper, we propose an activity classification method from visual lifelogs based on Fisher kernels, which extract discriminative embeddings from Hidden Markov Models (HMMs) of occurrences of semantic concepts. By using the gradients as features, the resulting classifiers can better distinguish different activities and from that we can make inferences about human behaviour. Experiments show the effectiveness of this method in improving classification accuracy, especially when the semantic concepts are initially detected with low degrees of accuracy.

Author Keywords
Visual lifelogging; everyday activities; human behaviour; concept attribute; Fisher kernel; HMM.

ACM Classification Keywords
H.3.3 [Information Storage and Retrieval]: Information Search and Retrieval

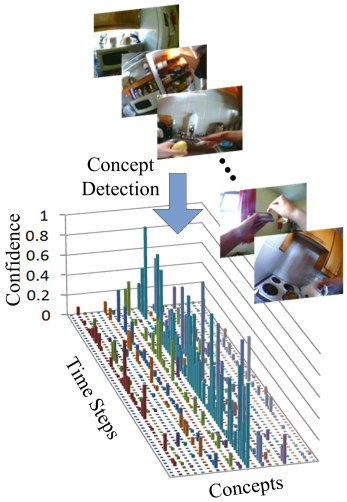

Figure 1: The dynamics of concept attributes quantified by confidences returned by concept detections.

Introduction

As an important source of quantified self information, lifelogging [5] is concerned with digitally capturing media-rich representations of everyday activities, making available a person's experiences in the form of events, interactions and relationships. Such rich pools of information are collected by individuals to characterize their own activities and behaviour for a variety of use cases. Visual recordings of events contain rich semantics which can be used to infer information about underlying activities including 'Who', 'What', 'Where' and 'When'. Visual lifelogging has now developed as an important aspect of the quantified self field which represents human behaviour using image-based or video-based media.

Whatever the reason for collecting such personal lifelog images or videos, be it for posterity, medical or well-being reasons, memory support or just for leisure, finding discrete human activities of interest from large lifelog collections and interpreting them via semantic meaning is crucial if we are to build real-world applications which focus on human behaviour. State-of-the-art approaches to identifying semantics from visual media use statistical techniques to map low-level local or global features like colour, texture or shape, to high-level semantic concepts like "indoor", "building" or "walk", a process termed "concept detection". The natural progression is from a lifelog image or video, to a set of such semantic concepts occurring in the image, and then to infer an activity the wearer was participating in while the image was taken based on the presence or absence of semantic concepts. Following that, we can then to aggregate activities over a long period of time in order to reason about the wearer's behaviour or identify changes in it.

Though effective in annotating visual media with individual concepts, concept detection in lifelogging has, to date, mostly failed to exploit temporal relationships among concepts which could provide useful information for activity or event classification. Some previous work has looked at temporal modeling of activities on top of concept detection in order to enhance the accuracy of either the underlying semantic concept detection [12], or the resulting activity characterization [11]. In [2], the authors showed that combining concept detection with temporal representations of those concept occurrences is promising when detecting more complex events on top of which events can then be inferred. In [8], Fisher kernel techniques were applied to encode the transitions between concept occurrences and absences over time. This encoding was into a compact and fixed-length feature vector which was used as the basis for further classification. Motivated by this previous work on modelling of events from a temporal viewpoint, we propose to apply Hidden Markov Models (HMMs) to model the time-varying dynamics of concept attributes, capturing the streams of occurrence and absense of semantic concepts individually and in combination. The Fisher scores are then extracted from the resulting generative model to form a set of even more compact and discriminate features. In theory this method has the capability of combining the advantages of generative and discriminative approaches in both temporal modeling and classification and we shall see how effective it is in practice.

Dynamics of Semantic Attributes

Using state-of-the-art concept detection methods, acceptable results can be achieved in some cases particularly for narrow domains and for concepts for which there exists enough annotated training data, according to the TRECVid benchmark [7]. The technology of automatic detection of concepts enables searching through visual lifelogs based on semantic attributes and this kind of content-based search on visual media has been validated as use-

ful for carrying out analysis of lifestyle behaviour patterns [3], [4]. However, despite recent progress, automatic concept detectors are still far from perfect and how to classify high-level events/activities based on such noisy semantic attributes needs to be tackled. This is especially important for cases where we then build upon the detected concepts such as using them to infer activities and then behaviour. For quantified self applications, there is a further challenge because of the diverse range of usable concepts, and the generally noisy nature of the lifelog data because of the wearers' movements and because even the images captured passively within the same lifelogged event may have significant perceptual differences.

Inspired by recent work on attribute-based temporal modeling [8], [11], we model the dynamic evolution of human activities using concept detection results as input. In effect this means that streams of activities are represented as sequences of units such as clips or frames. Concept detections are applied to each of units and by concatenating the output results (confidences) of pre-trained concept detectors at the same timestep as a vector, one activity stream can be represented by a temporally ordered sequence of vectors, as shown in Fig. 1. In a set of activity samples X, each activity can be structured as $X_n = \{x_{n1}, x_{n2}, ..., x_{nt}, ...\}$ with $x_{nt} \in \mathcal{R}^d$ representing a fixed length confidence vector whose dimension is equal to d, the number of concepts detected.

Using a HMM Fisher Kernel for Activity Classification

Since HMMs have previously been validated as effective in characterising lifelogging activities [11], we employed HMMs to encode the dynamic distributions of concepts throughout lifelog events. Assume there are l hidden states in the HMM and each pair of states have a transition probability $a_{ij} = P(s_i|s_j)$. The parameters of the HMM can be denoted as $\lambda = (A, B, \pi)$, where $A = \{a_{ij}\}$, $\pi = \{\pi_i\}$ stands for the initial state distribution. $b_j(X_t)$ is the distribution of the concept observation X_t at time step t with respect to state j. Because the confidence vector X_t has continuous values, we employed Gaussian emission distributions $b_j(X_t) = \mathcal{N}(X_t, \mu_i, \sigma_i)$ and $B = \{\mu_i, \sigma_i\}$. Parameters μ_i and σ_i are the mean and covariance matrix of the Gaussian distribution in state i respectively.

The principle of the Fisher kernel is that similar samples should have similar dependence on the generative model, i.e. the gradients of the parameters [6]. Instead of directly using the output of generative models, using a Fisher kernel tries to generate a feature vector which describes how the parameters of the activity model should be modified in order to adapt to different samples. Based on the above formalization of a HMM, X can be characterized as Fisher scores with regard to the parameters λ:

$$U_X = \nabla_\lambda logP(X|\lambda) = [\frac{\partial logP}{\partial a_{ij}}, \frac{\partial logP}{\partial u_{ik}}, \frac{\partial logP}{\partial \sigma_{ik}}, \frac{\partial logP}{\partial \pi_i}]^T \quad (1)$$

where $1 \leq i \leq l$ and $1 \leq k \leq d$. Therefore, the Fisher kernel can be formalized as $K(X_i, X_j) = U_{X_i}^T I_F U_{X_j}^T$, where $I_F = E_X(U_X U_X^T)$ denotes the Fisher information matrix.

Experiments and Evaluation

Experimental Setup

To evaluate our proposed activity classification algorithm, we performed a comprehensive assessment of our approach using datasets with various accuracies for semantic concept detection [11]. The 16 everyday activity types listed in Table 1 are used in the evaluation for which 10,497 lifelogged images have been collected from the SenseCam

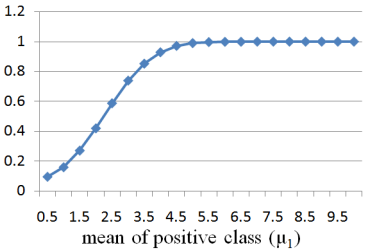

Figure 2: Averaged concept *MAP* with different μ_1 values.

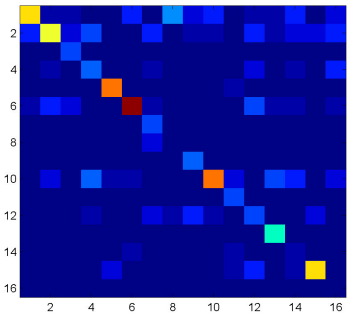

Figure 3: Confusion matrix when using HMM log-likelihood representation as features.

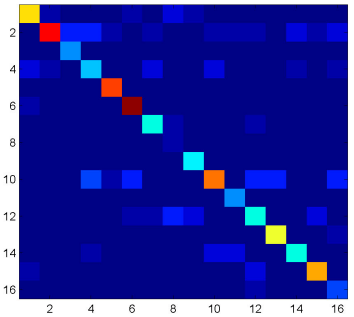

Figure 4: Confusion matrix for our proposed Fisher kernel-based classification.

wearable camera, worn by 4 people with different demographics. SenseCam[1] is a camera, worn around the neck and facing forward, which continuously captures images from a first-person view of the wearer. Note that the activities in Table 1 are chosen based on time dominance, generality and high frequency of occurrence [10], and can be used to support applications like independent living assistance, obesity analysis, and chronic disease diagnosis.

Eating	Drinking	Cooking
Clean/Tidy	Use computer	Watch TV
Child care	Food shopping	Shopping (non-food)
Reading	Using phone	Driving
Taking bus	Walking	(listen to) Presentation
Talking		

Table 1: Everyday activity types in our evaluation.

Since we wish to explore the impact of the accuracy of semantic concept detection as a variable in the recognition of activities, concept detection results are simulated based on groundtruth annotations, following the work of [1] and [11]. By simulating downgrading of the detection accuracy based on a 100% accurate ground truth as a starting point we can control the levels of concept detection accuracies and in this way a comprehensive comparison can be carried out to evaluate our proposed activity classification in a realistic setting. In this experiment, concept detectors for 85 concepts are simulated by changing the controlling parameter μ_1 [11] and Fig. 2 shows averaged concept MAP (mean average precision) by 20 simulation runs.

Baselines

In order to evaluate the performance of the Fisher kernel in a discriminative classifier, we employed two widely used classifiers: support vector machines (SVMs) and k-nearest neighbor classifiers (KNNs).

The generative method based on HMMs as used in [11] is employed as one baseline. The HMMs are first trained for each activity class and we concatenate the log-likelihood representations of per-class posteriors into a vector. The LibSVM [2] implementation of SVMs with the linear kernel is employed to perform SVM classifications on log-likelihood representations.

For the KNN classifier, dynamic time warping (DTW) is applied to the activity samples based on Euclidian similarity, i.e. minimizing the sum of distances between corresponding samples. As previously pointed out, the length of the human activity will naturally vary across different classes or samples and this step is to perform temporal alignment on these variable-length time series.

Results

To alleviate the sub-optimal problem of Fisher kernels induced by (nearly) zero gradient representations of a generative model, we employed a model parameter learning as proposed in [9], to train the model so that samples of the same class will have more similar gradients than the other classes. The Fisher kernel is then embedded in the SVMs for activity classification. To simplify the computation, we approximate I_F by the identity matrix in the implementation.

In Table 2, activity classification accuracies are listed for all methods at different performance levels for concept detection. For classifications based on log-likelihood (HMM+SVM) and on Fisher kernel (FK+SVM), the generative models are obtained with two-state ergodic HMMs to model the sequence of concept occurrences. Following [11],

[1] www.research.microsoft.com/sensecam/

[2] http://www.csie.ntu.edu.tw/∼cjlin/libsvm/

	Concept Detection Performances (*MAP*s)			
Methods	*0.095*	*0.157*	*0.265*	*0.412*
DWT+KNN	10.0 ± 1.7	20.8 ± 2.8	40.1 ± 6.4	59.1 ± 2.6
HMM+SVM	26.2 ± 2.2	65.2 ± 1.6	69.8 ± 1.8	77.4 ± 2.1
FK+SVM	50.0 ± 2.8	72.7 ± 2.2	80.6 ± 2.6	85.1 ± 1.5
	0.580	*0.731*	*0.848*	*0.925*
DWT+KNN	73.5 ± 4.8	81.3 ± 2.8	85.8 ± 1.7	89.1 ± 0.8
HMM+SVM	82.1 ± 3.1	85.1 ± 1.6	85.9 ± 1.5	86.6 ± 1.2
FK+SVM	86.1 ± 2.6	87.6 ± 0.3	90.2 ± 2.0	89.2 ± 0.8

Table 2: Accuracy comparison (in percentages) at different concept detection accuracies.

	Hidden States		
***MAP*s**	*5*	*10*	*20*
0.095	0.50	0.50	0.50
0.157	0.72	0.73	0.75
0.265	0.81	0.82	0.84
0.412	0.87	0.88	0.89
0.580	0.90	0.91	0.93
0.731	0.92	0.93	0.95
0.848	0.93	0.95	0.96
0.925	0.94	0.95	0.97

Table 3: Performances of FK+SVM with different numbers of hidden states.

we applied latent semantic analysis to map the original attribute space to a more compact 35-dimensional space according to the contextual correlation of concept occurrences. Multivariate Gaussian emission probabilities with full covariance matrices are employed in the HMMs to model the high dimensional features.

As shown in Table 2, the classification based on the Fisher kernel significantly out-performs the baselines across various concept detection accuracies. Most especially, when concept detection accuracies are not high, such as when the $MAP < 0.5$ which is a realistic expectation according to the TREVid benchmark [7], the improvement can be as high as greater than 10% for most cases. This suggests that our proposed method can encode the dynamic features of concept occurrences into more discriminative features. This is especially useful for real-world quantified self applications where concept detections are noisy and inaccurate due to reasons including visual diversity, user movement of the camera causing blurring, image quality, etc. The discriminative capability of the proposed method is also illustrated by Fig. 3 and 4 in which the two features extracted from HMMs are used at the same concept detection MAP (0.157). While more samples are mistakenly classified using HMM log-likelihood representations (average accuracy 65.2%), less mis-classifications are made when embedding SVMs with Fisher scores and kernels (average accuracy 72.7%).

In addition to the performances demonstrated in Table 2 using two hidden states, the results for the proposed FK+SVM methods with 5, 10, and 20 states are also shown in Table 3. As shown in the table, similar results to Table 2 are obtained across different numbers of hidden states. This reflects the robustness of the Fisher kernel-based activity classification method.

Conclusions

Automatic detection of human activities based on visual lifelogs represents a natural use of such lifelogs. However, due to the variety of activities humans are involved in combined with the movement of the wearable lifelog camera as images are taken, the diversity of concepts and the quality of images poses challenges to reliably detecing human activities form visual media. This is partly due to the difficulty of discriminating activities of interest from others. In this paper, we propose to employ a Fisher kernel to extract embedding from HMMs which model human activities, for more accurate activity classification based on concept occurrences. Experimental results have shown the advantage of reflecting temporal features and making classification more accurate, especially when concept detection has poor performance as measure by MAP, which is common in real-world quantified self applications.

Acknowledgements

This work was funded by the National Natural Science Foundation of China under Grant No. 61272231, 61472204, Beijing Key Laboratory of Networked Multimedia and by Science Foundation Ireland under grant SFI/12/RC/2289.

REFERENCES

1. Robin Aly, Djoerd Hiemstra, Franciska de Jong, and Peter Apers. 2011. Simulating the future of concept-based video retrieval under improved detector performance. *Multimedia Tools and Applications* (2011), 1–29.
2. Subhabrata Bhattacharya, Mahdi Kalayeh, Rahul Sukthankar, and Mubarak Shah. 2014. Recognition of Complex Events exploiting Temporal Dynamics between Underlying Concepts. In *CVPR 2014*. 2243–2250.
3. Daragh Byrne, AidenR. Doherty, CeesG.M. Snoek, GarethJ.F. Jones, and AlanF. Smeaton. 2010. Everyday concept detection in visual lifelogs: validation, relationships and trends. *Multimedia Tools and Applications* 49, 1 (2010), 119–144. DOI: http://dx.doi.org/10.1007/s11042-009-0403-8
4. Aiden R. Doherty, Niamh Caprani, Ciaran O'Conaire, Vaiva Kalnikaite, Cathal Gurrin, Noel E. O'Connor, and Alan F. Smeaton. 2011. Passively recognising human activities through lifelogging. *Computers in Human Behavior* 27, 5 (2011), 1948–1958.
5. Cathal Gurrin, Alan F. Smeaton, and Aiden R. Doherty. 2014. LifeLogging: Personal Big Data. *Foundations and Trends in Information Retrieval* 8, 1 (2014), 1–125. DOI: http://dx.doi.org/10.1561/1500000033
6. Tommi S. Jaakkola and David Haussler. 1999. Exploiting Generative Models in Discriminative Classifiers. In *NIPS 1999*. 487–493.
7. Alan F. Smeaton, P. Over, and W. Kraaij. 2008. High Level Feature Detection from Video in TRECVid: a 5-year Retrospective of Achievements. In *Ajay Divakaran (Ed.), Multimedia Content Analysis, Theory and Applications*. Springer, 151–174.
8. Chen Sun and Ram Nevatia. 2013. ACTIVE: Activity Concept Transitions in Video Event Classification. In *ICCV 2013*. 913–920.
9. Laurens van der Maaten. 2011. Learning Discriminative Fisher Kernels. In *ICML 2011*. 217–224.
10. P. Wang and A.F. Smeaton. 2012. Semantics-based selection of everyday concepts in visual lifelogging. *International Journal of Multimedia Information Retrieval* 1, 2 (2012), 87–101.
11. Peng Wang and Alan F. Smeaton. 2013. Using visual lifelogs to automatically characterize everyday activities. *Information Sciences* 230, 0 (2013), 147–161. DOI: http://dx.doi.org/10.1016/j.ins.2012.12.028
12. Peng Wang, Alan F. Smeaton, and Cathal Gurrin. 2015. Factorizing Time-Aware Multi-way Tensors for Enhancing Semantic Wearable Sensing. In *MMM 2015*. 571–582.