ENHANCING THE DETECTION OF CONCEPTS FOR VISUAL LIFELOGS USING CONTEXTS INSTEAD OF ONTOLOGIES

Peng Wang[†]*, Alan F. Smeaton*[‡]*, Yuchao Zhang*[†]*, Bo Deng*[†]

[†]Beijing Institute of System Engineering, Beijing, 100101, P. R. China
[‡]Insight Centre for Data Analytics, Dublin City University, Glasnevin, Dublin 9, Ireland
{pwang, asmeaton}@computing.dcu.ie, dragonzyc@163.com, bodeng@vip.tom.com

ABSTRACT

Automatic detection of semantic concepts in visual media is typically achieved by an automatic mapping from low-level features to higher level semantics and progress in automatic detection within narrow domains has now reached a satisfactory performance level. In visual lifelogging, part of the quantified-self movement, wearable cameras can automatically record most aspects of daily living. The resulting images have a diversity of everyday concepts which severely degrades the performance of concept detection. In this paper, we present an algorithm based on non-negative matrix refactorization which exploits inherent relationships between everyday concepts in domains where context is more prevalent, such as lifelogging. Results for initial concept detection are factorized and adjusted according to their patterns of appearance, and absence. In comparison to using an ontology to enhance concept detection, we use underlying contextual semantics to improve overall detection performance. Results are demonstrated in experiments to show the efficacy of our algorithm.

Index Terms— Visual lifelogging, concept detection, non-negative matrix factorization, concept semantics.

1. INTRODUCTION

Lifelogging is the term used to describe the process of automatically, and ambiently, digitally recording our own day-to-day activities for our own personal purposes [1]. With the proliferation of mobile devices with their computational capability, lightweight nature and long-life battery, research on applying lifelogging techniques across several domains has become more feasible. *Visual lifelogging* is the term used to describe one class of personal lifelogging which employs wearable cameras to capture image or video of everyday activities. These include SenseCam [2], Vicon Revue, as well as the possibilities offered by Google Glass.

Many projects now use visual lifelogging in applications like aiding human memory, diet monitoring, chronic disease diagnosis, recording activities of daily living (ADL) and so on. Microsoft Research have pioneered this with the development of the SenseCam and there is evidence that SenseCam images can improve memory recall for people with memory disorders [3]. Though dietary patterns are proven as a critical contributing factor to many chronic diseases, traditional strategies based on self-reported information do not fulfill the task of accurate diet reporting. DietSense [4] is an example of lifelogging using mobile devices to support automatic multimedia documentation of dietary choices. More recently, evidence suggests that visual lifelogs provide a more accurate measure of energy intake while individuals' self-report often underestimate the true value [5]. IMMED [6] is a typical application of visual lifelogging of ADL, in which video data of the instrumented activities of a patient are recorded and indexed to assess the cognitive decline caused by dementia.

Metadata like date, time and location may be sufficient for many lifelogging applications but there are others which require searching through lifelogs based on content, and for this to happen the automatic detection of semantic concepts needs to be introduced. Visual lifelogs represent a new form of multimedia which require semantic indexing and retrieval, for which much preliminary work has already been done in other domains. State-of-the-art techniques use statistical approaches to map low level image features to concepts which can then be fused. According to the TRECVid benchmark [7], acceptable results have been achieved already in many cases particularly for concepts for which there exists enough annotated training data. Individual concepts detected by standalone classifiers can be fused to determine high-level semantics though this demands a high level of classification accuracy for the underlying concepts [8]. However, in the visual lifelogging domain, the challenge of improving detection accuracy is more severe given the visual diversity of lifelog content and the large variety of concepts compared to, say, broadcast TV news. Even the images captured passively within the same lifelogged event may have significant perceptual differences due to the wearers' movements. Furthermore, the one-per-class SVM classifiers which are widespread in analysis of other kinds of images and video, will ignore whatever

This work was supported by Science Foundation Ireland under grant SFI/12/RC/2289.

concept relationships may exist.

In this paper, we proposed an approach to model everyday concept occurrence patterns using concept detection results as the only input, which we use to improve concept detection in visual lifelogs. This exempts us from the overhead of building an ontology and the need for training data. In order to evaluate the effectiveness of this methodology, we employed SenseCam (shown in Figure 1) as a wearable device to log details of users' lives. SenseCam is a lightweight passive camera with several built-in sensors which captures the view of the wearer with its fisheye lens. By default, images are taken at the rate of about one every 50 seconds while the on-board sensors can help to trigger the capture of pictures when sudden changes are detected in the environment of the wearer. Some typical events from SenseCam images are shown in Figure 2.

Fig. 1. A variety of wearable visual lifelog devices through the ages including SenseCam (bottom right).

The contribution of our work is three-fold. We present an algorithm based on non-negative matrix factorization to improve concept detection accuracy and we then model everyday concept semantics to learn the appearance patterns of those concepts which have low-accuracy detection, which we then boost. Finally, we compare our approach with an ontology-based method for improving concept detection and we demonstrate this to be advantageous in efficacy, implementation and domain independence.

The rest of the paper is organized as follows: in Section 2 we describe the multi-concept problem and we introduce contextual semantics in lifelogged media. Our algorithm for enhancing the detection of everyday concepts is discussed in Section 3 and the experimental implementation and results analysis are presented in Section 4. Finally, we close the paper with conclusion and future work.

2. CONTEXTUAL SEMANTICS OF EVERYDAY CONCEPTS

Concepts express the semantics of media in a useful way and are usually detected by providing a meaningful link between low-level features, and high-level understanding.

2.1. Confidence-based Concept Detection

Following the state-of-the-art in concept detection, we employ a generic one-per-class SVM classifier with confidence results as a basis for our algorithm. This means each concept detector is responsible for performing detection of a single concept but with a distance between each instance and a hyper plane which exists between positive and negative concept instances.

By defining a target concept of an image x as c, the distance of image x belonging to concept c returned by the classifier is represented as $d_c(x)$, whose sign and magnitude reflect the class prediction and the confidence level of that prediction. Since the returned distance is uncalibrated, it can not be used as confidence for concept judgement. A widely employed method is to use a sigmoid function to fit the posterior probability of $P(c|x)$ [9]:

$$Conf(c|x) = \frac{1}{1 + exp(Ax + B)} \quad (1)$$

where parameters A and B can be obtained by fitting the maximum likelihood estimation from the training dataset. Assume we have M classifiers for M concepts, we directly binarize $Conf(c|x)$ to obtain the appearance of each concept in image x.

2.2. Exploiting Everyday Concept Semantics

Since concepts usually co-occur within even a single image rather than in isolation, as we can see from the exemplar images in Figure 2, the understanding of lifelogged media such as SenseCam image streams is actually a multi-concept detection problem. While we use one-per-class classifiers, intrinsic relationships between concepts are neglected and ultimately this ends up with multiple isolated binary classifiers which do not exploit concept semantics. This approach is likely to suffer from the shortcomings of misclassification or inconsistency among the detected concepts.

It is widely accepted that there is strong correlation among concept ontological semantics and contextual semantics. For example, 'Road' can be modeled as a subclass of 'Outdoor' while 'Indoor' and 'Outdoor' are usually mutually disjoint. This is consistent with some cases of activities like 'Driving', 'Walking' and so on, in which the concepts 'Outdoor' and 'Road' co-occur more frequently while 'Indoor' and its descendant concepts like 'Office', 'Kitchen','Computer', etc. are excluded and have less likelihood to appear in such activities. While the construction of an ontology is domain-specific and usually subjective, the use of contextual semantics has great potential, as pointed out more recently in the lifelogging domain by [10]. Our algorithm exploiting such contextual semantics to enhance everyday concept detection, is now described.

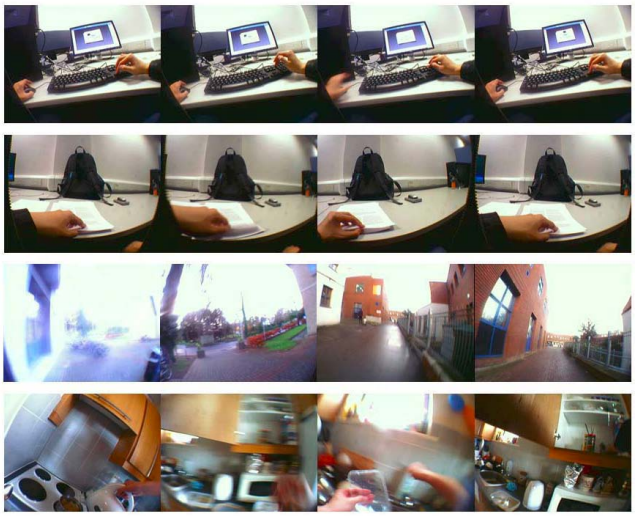


Fig. 2. Samples of lifelogging activities (top to bottom: 'Using computer', 'Reading', 'Walking', 'Cooking', for each row).

3. EVERYDAY CONCEPT DETECTION ENHANCEMENT

Our algorithm is based on the concepts detection results from a series of images taken from events which are automatically segmented based on the technique introduced in [11]. An event corresponds to a single activity in the wearer's day such as watching TV, commuting to work, or eating a meal, with an average stream of 20 events of varying duration in a typical day.

3.1. Problem Formalization

We assume a universe of concepts C. Let $\{E_1, E_2, ..., E_n\}$ be the set of event streams in the dataset. Event E_i is represented by successive images $I^{(i)} = \{Im_1^{(i)}, Im_2^{(i)}, ..., Im_k^{(i)}\}$. Each image $Im_j^{(i)}$ might have several concepts detected, we assume the concepts appearing in image $Im_j^{(i)}$ are represented as a confidence vector $C_j^{(i)} = \{c_{j1}^{(i)}, c_{j2}^{(i)} ... c_{jM}^{(i)}\}$ for M concepts. The whole set of SenseCam images can be denoted as $I = \{I^{(1)}, I^{(2)}, ..., I^{(n)}\}$ which has dimension $N = \sum_{i=1}^{n} k_i$, where k_i is the number of images in each event E_i. Concept detection for these N images for M concepts can be described as a confidence matrix:

$$\mathbf{C} = \begin{pmatrix} c_{11} & c_{12} & \dots & c_{1M} \\ c_{21} & c_{22} & \dots & c_{2M} \\ \vdots & \vdots & \ddots & \vdots \\ c_{N1} & c_{N2} & \dots & c_{NM} \end{pmatrix} \quad (2)$$

The task now is to modify the $N \times M$ dimensional matrix C in order to keep consistency with the underlying contextual pattern of concepts. According to [12], matrix C can be represented as $C \approx UH$, in which r columns of U are basis and each column of H is an encoding in one-to-one correspondence with a column in C. The intuition of this is to form the confidence matrix by simply combining partial information (columns in U) with an additive operator since all elements in H are non-negative. That is to say, various concepts can be mapped to combinations of semantic units and concept-concept contextual semantics can be evaluated through this new sparse encoding.

3.2. Factorizing the Detection Results

Let the dimensions of component matrix U and H be $N \times r$ and $r \times M$. Since r is the reduced rank satisfying $(N+M)r < NM$, the approximation of UH is indeed the compression of C. The approximation factorization defined above can be solved by optimizing the cost function defined to qualify the quality of the approximation. Different forms of cost function and corresponding optimization can be applicable to this problem but in factorizing the confidence matrix, the weighted measure is more suitable since detection performance is different due to the characteristics of concepts and quality of the training set. To distinguish the contribution of different concept detectors to the cost function, the weighted cost function is employed as

$$F = \frac{1}{2}\|W \circ (C - UH)\|_F^2 = \frac{1}{2}\sum_{ij} w_{ij}(c_{ij} - U_{i\cdot}H_{\cdot j})^2 \quad (3)$$

such that $U \geq 0, H \geq 0$, where $\circ$ denotes element-wise multiplication, $W = (w_{ij})_{N \times M}$ denotes the weight matrix and $\|\cdot\|_F^2$ denotes the Frobenius norm. Gradient descent method can be applied for optimizing this problem, implemented by updating U and H in the opposite direction to the gradient at each iteration through

$$U = U - \alpha_U \partial F / \partial U \quad (4)$$

$$H = H - \alpha_H \partial F / \partial H \quad (5)$$

after each step α_U, α_H. $\partial F/\partial U$ and $\partial F/\partial H$ can be calculated by

$$\frac{\partial F}{\partial U} = [(UH - C) \circ W]H^T \quad (6)$$

$$\frac{\partial F}{\partial H} = U^T[(UH - C) \circ W] \quad (7)$$

and we employed α_U, α_H as the form

$$\alpha_U = U/[(UH \circ W)H^T] \quad (8)$$

$$\alpha_H = H/[U^T(UH \circ W)] \quad (9)$$

where / denotes element-wise division. Note that it is not hard to prove that under such updating rules, the cost function in Equation 3 is non-increasing in each optimization step.

3.3. Concept Detection Enhancement

To obtain a reconstruction of the underlying semantic structure, the weights need to be set in terms of concept accuracy. Because each confidence value c_{ij} in C denotes the probability of concept c_j occurring in the image, estimating the existence of c_j is more likely to be correct when c_{ij} is high enough. Under this premise, we carried out the concept detection enhancement as follows.

First, each column of confidence matrix C is normalized at $Max - Min$ scale. This is then followed by constructing a new sparse matrix C' by thresholding C, whose element is $c'_{ij} = c_{ij}$ if $c_{ij} \geq threshold$ or 0 otherwise. The rationale for this is to retain elements with high confidence as "seeds" and apply the contextual information modeled by non-negative factorization to predict other concepts in correlation with these seed concepts. A sparse confidence matrix C' is achieved and we denote the non-zero element set in C' as C_1. Meanwhile, the set of $C' - C_1$ can be used to denote zero elements which need to be estimated from C_1.

Then C' is factorized using the algorithm described in Section 3.2. This involves the iterative optimization of the cost function defined in Equation 3. In the optimization step, we configure the settings of weights as $w_{ij} = 1$ if $c'_{ij} \in C_1$, otherwise $w_{ij} \in (0, 1)$. In this step, two factor matrix of U and H are returned as an estimation of the contextual structure of C'.

Finally, the approximation of C' can be calculated as $c'_{ij} = \sum_{k=1}^{r} u_{ik} h_{kj}$. The new confidence values for elements in $C' - C_1$ can form an estimate of concept detection to adjust the original detection result by averaging the original confidence and the new estimated value.

4. EXPERIMENTS AND EVALUATION

4.1. Experiment Setup and Dataset

To assess the performance of our algorithm, we used a set of 85 everyday concepts as investigated in [13]. We used a dataset including event samples of 23 activity types collected from 4 SenseCam wearers and consisting of 12,248 SenseCam images [1]. Concept detectors with different accuracy levels were applied and the metrics of AP and MAP were calculated for concepts based on manual groundtruth. Different concept detection accuracies were provided in the dataset by varying the mean of positive class μ_1 in the range [0.5...10.0]. For each setting of parameters, we executed 20 repeated runs to avoid random performance and the averaged concept AP and MAP were both calculated. Figure 3 shows the improvement in concept MAP with increasing μ_1 and near-perfect detection performances are achieved when $\mu_1 \geq 5.5$.

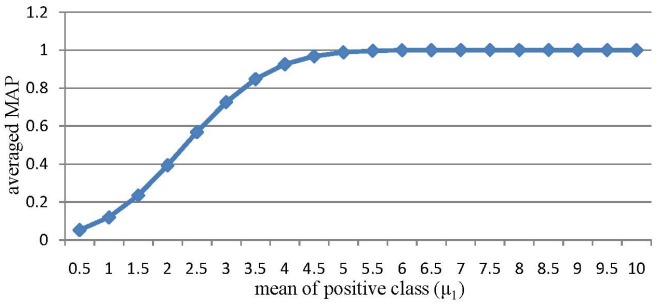


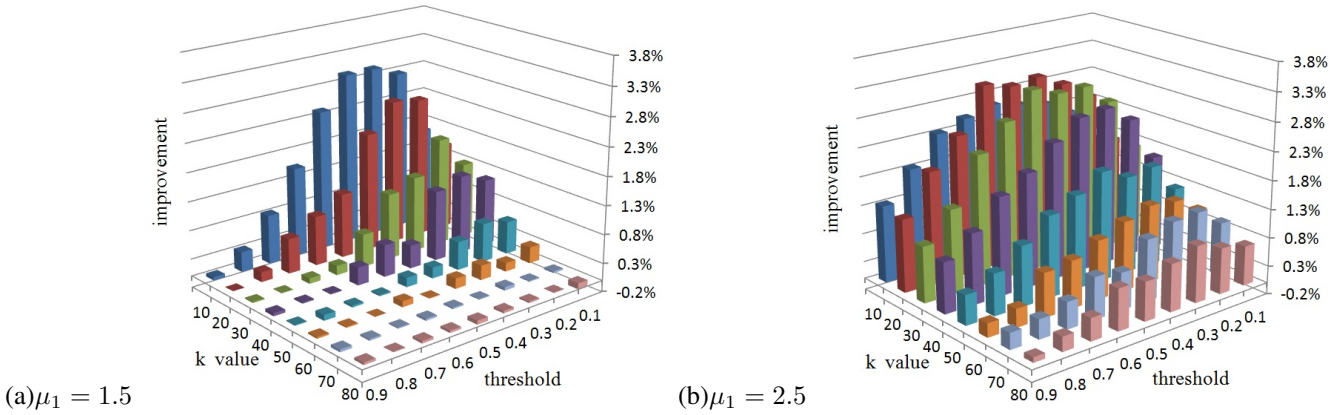
Fig. 3. Averaged concept *MAP* with different μ_1 values.

Both contextual and ontological methods are implemented and compared in our experiments. Contextual enhancement is carried out as described in Section 3 with concept detection confidence as the only input. In implementing an ontology-based adjustment algorithm, an everyday concept ontology was first constructed for the set of 85 concepts using the ontology language OWL, which is a standard Semantic Web language. Both *subsumption* and *disjointness* concept relationships are used. Subsumption is a relationship restricting the membership of a concept. By relating two concepts with disjointness, no instance of either class can be an instance of both classes. In our implementation, the state-of-the-art Semantic Web reasoner is also embedded straightforwardly to leverage explicit statements in the ontology to create logically valid but implicit statements. Since the ontological method has to learn the correlation of accuracy and multi-concept confidences before enhancement, we randomly select half the dataset for training and the other half for evaluation. The sigmoid function is used for fitting the correlation which has a form similar to Equation 1 except for the setting of parameters.

4.2. Contextual vs. Ontological

The effectiveness of contextual enhancement is demonstrated in Figure 4, in which improvement is depicted at two different concept detection accuracies, determined by $\mu_1 = 1.5$ and $\mu_1 = 2.5$ respectively. Instead of choosing a particular setting of parameters, results for $k \in [10, ..., 80]$ and $threshold \in [0.1, ..., 0.9]$ are shown in Figure 4. All cases in Figure 4 are achieved by executing the algorithm in 20 runs and the averaged MAP improvement across all 85 concepts are obtained.

As shown in Figure 4, the performance of contextual enhancement is better when the positive class mean increases from 1.5 to 2.5. While there are cases in which overall concept detection is not obviously improved in Figure 4(a), detection performance is improved in all cases shown in Fig-



(a)$\mu_1 = 1.5$ (b)$\mu_1 = 2.5$

Fig. 4. MAP improvement with various parameter configures.

ure 4(b). This is consistent with the premise that there are some concepts whose detections are satisfactory and can be selected as seeds to boost the performance of the other concept detectors. In Figure 4(a), when $threshold$ is chosen as too high, there will be fewer correct concept detection results chosen, hence overall performance could hardly be improved. The situation is alleviated in Figure 4(b), in which the original detection performance is better, as shown in Figure 3. That means in Figure 4(b), more correctly detected concepts can be used to give better estimations on the others, based on contextual semantics modeled by the algorithm. The choice of too "noisy" concepts can also degrade the improvement, as depicted when $threshold$ is small. In these cases, erroneous detection results are likely to be chosen to C_1 in the thresholding procedure as described in Section 3.3. The best overall performances are achieved when $threshold$ value is around 0.4.

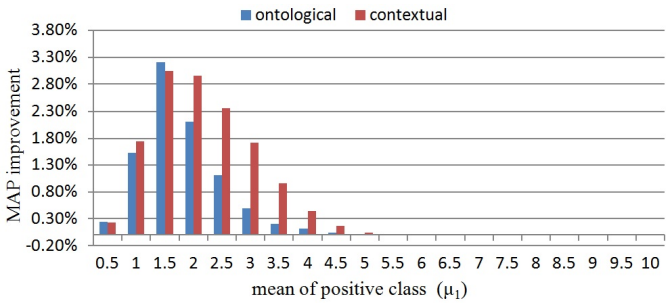


Fig. 5. Improvement comparison of contextal and ontological approaches.

Since averaging MAP over different detection accuracies is meaningless, pairwise comparison is depicted in Figure 5 at different μ_1 values where contextual enhancement ($k = 10$, $threshold = 0.3$) significantly outperforms the ontological approach in most cases, except for $\mu_1 = 0.5$ and $\mu_1 = 1.5$. This is because the ontological approach uses one half of the dataset to learn the distribution and only half of the dataset for testing. Based on the prior knowledge deliberately learned from extra training data, the ontological approach adapts to the distribution of the dataset more easily. However, no extra training data or processing are needed in our algorithm, which can self-learn the contextual semantics of concepts and enhance overall detection performance. The poor performance of both approaches at $\mu_1 = 0.5$ makes sense as the initial detection accuracy is just too low. As shown in Figure 3, the overall MAP at $\mu_1 = 0.5$ is nearly zero. In this case, no correctly detected concept can be selected and utilised which is impractical in real world applications. When initial detection performance is good enough, as shown in Figure 3 if $\mu_1 \geq 5.5$, there is no space to improve detection accuracy. Therefore, for both approaches in Figure 5, the improvement is not that significant at $\mu_1 \geq 5.5$.

Concept coverage is another advantage of our algorithm as demonstrated in Figure 6, using the same parameters as Figure 5. In Figure 6, the peak performances of contextual and ontological approaches (at $\mu_1 = 1.5$) are visualized across all 85 concept APs. Nearly 80 concepts are improved by our algorithm whereas the number of improved concepts by ontological approach is only 30. Because the ontological approach is based on a pre-constructed ontology and a set of training data, it is constrained by ontological concept relationships and having enough positive samples. In our experiment, 52 concepts have available parent and disjoint concepts after inference, and 35 of these have more than 100 positive samples for distribution training. It seems that the ontological approach outperforms our algorithm at limited cases, say, the first 5 concepts in Figure 6. However, this is because of the prior knowledge learned from one half of the dataset and training on the other half, which is indeed a limitation of the ontological approach. On the contrary, our algorithm is effective for most of the concepts and the overall improvement is significant across all 85 concepts in a two-way ANOVA test with 20 replications ($p < 0.01$). Similar results can be obtained using other parameter settings.

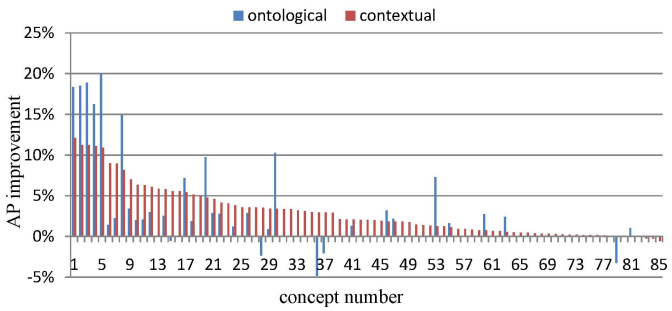


Fig. 6. Improvement comparison over all concepts.

According to the above results, our contextual enhancement algorithm has many advantages. First, it is data efficient and easy to implement since no prior knowledge is needed unlike ontology construction or distributions learned from extra training data. Second, it is shown to be effective in significantly improving detection accuracies for a large number of concepts. Finally, the only input is initial concept detection results and the algorithm is independent of any specific implementation of concept detectors.

5. CONCLUSIONS AND FUTURE WORK

An enhancement algorithm is described for improving everyday concept detection performance in visual lifelogging. Based on non-negative matrix factorization, the algorithm can model global contextual semantics through partial concept detection results which have better accuracies. The confidences of less accurate concept detections are estimated and combined with the initial results to enhance the overall detection performance. By comparison with an ontology-based approach, experiments demonstrate that our algorithm has advantages in many aspects such as easy implementation, effectiveness, high concept coverage and domain independence. Our future work is to apply this method to learn pairwise concept correlations and exploit the automatic construction of contextual semantic networks for visual lifelogging.

Acknowledgement: The research reported in this paper was funded by Science Foundation Ireland under grant S-FI/12/RC/2289.

6. REFERENCES

[1] P. Wang and A.F. Smeaton, "Using visual lifelogs to automatically characterise everyday activities," *Information Sciences*, vol. 230, pp. 147–161, 2013.

[2] S Hodges, L Williams, E Berry, S Izadi, J Srinivasan, A Butler, G Smyth, N Kapur, and K Wood, "SenseCam: A retrospective memory aid," in *UbiComp 2006: Ubiquitous Computing*, pp. 177–193. Springer, 2006.

[3] G Browne, E Berry, N Kapur, S Hodges, G Smyth, P Watson, and K Wood, "Sensecam improves memory for recent events and quality of life in a patient with memory retrieval difficulties," *Memory*, vol. 19, no. 7, pp. 713–722, 2011.

[4] S. Reddy, A. Parker, J. Hyman, J. Burke, D. Estrin, and M. Hansen, "Image browsing, processing, and clustering for participatory sensing: lessons from a dietsense prototype," in *Proceedings of the 4th workshop on Embedded networked sensors*. ACM, 2007, pp. 13–17.

[5] L. Gemming, A.R. Doherty, P. Kelly, J. Utter, and C. Ni Mhurchu, "Feasibility of a sensecam-assisted 24-h recall to reduce under-reporting of energy intake," *European journal of Clinical Nutrition*, vol. 67, no. 10, pp. 1095–1099, 2013.

[6] R. Mégret, V. Dovgalecs, H. Wannous, S. Karaman, J. Benois-Pineau, E. El Khoury, J. Pinquier, P. Joly, R. André-Obrecht, Y. Gaëstel, and J.-F. Dartigues, "The IMMED project: wearable video monitoring of people with age dementia," in *Proceedings of the international conference on Multimedia*, 2010, pp. 1299–1302.

[7] A.F. Smeaton, P. Over, and W. Kraaij, "High level feature detection from video in TRECVid: a 5-year retrospective of achievements," in *Ajay Divakaran (Ed.), Multimedia Content Analysis, Theory and Applications*. 2008, pp. 151–174, Springer.

[8] P. Toharia, O. Robles, A.F. Smeaton, and Á. Rodríguez, "Measuring the influence of concept detection on video retrieval," in *Computer Analysis of Images and Patterns*. Springer, 2009, pp. 581–589.

[9] J.C. Platt, "Probabilistic outputs for support vector machines and comparison to regularized likelihood methods," in *Advances in Large Margin Classifiers*, 1999.

[10] D. Byrne, A.R. Doherty, Cees G. M. Snoek, G. Jones, and A.F. Smeaton, "Everyday concept detection in visual lifelogs: validation, relationships and trends," *Multimedia Tools Appl.*, vol. 49, no. 1, pp. 119–144, 2010.

[11] H. Lee, A.F. Smeaton, N. O'Connor, G. Jones, M. Blighe, D. Byrne, A. Doherty, and C. Gurrin, "Constructing a SenseCam visual diary as a media process," *Multimedia Syst.*, vol. 14, no. 6, pp. 341–349, 2008.

[12] D.D. Lee and H.S. Seung, "Learning the parts of objects by nonnegative matrix factorization," *Nature*, vol. 401, pp. 788–791, 1999.

[13] P. Wang and A.F. Smeaton, "Semantics-based selection of everyday concepts in visual lifelogging," *International Journal of Multimedia Information Retrieval*, vol. 1, no. 2, pp. 87–101, 2012.