Semantically Smoothed Refinement for Everyday Concept Indexing

Peng Wang *, Lifeng Sun, Shiqiang Yang, and Alan F. Smeaton

National Laboratory for Information Science and Technology
Department of Computer Science and Technology, Tsinghua University, P. R. China
pwang@tsinghua.edu.cn, sunlf@tsinghua.edu.cn, yangshq@tsinghua.edu.cn
Insight Centre for Data Analytics
Dublin City University, Glasnevin, Dublin 9, Ireland
alan.smeaton@dcu.ie

Abstract. Instead of occurring independently, semantic concepts pairs tend to co-occur within a single image and it is intuitive that concept detection accuracy for visual concepts can be enhanced if concept correlation can be leveraged in some way. In everyday concept detection for visual lifelogging using wearable cameras to automatically record everyday activities, the captured images usually have a diversity of concepts which challenges the performance of concept detection. In this paper a semantically smoothed refinement algorithm is proposed using concept correlations which exploit topic-related concept relationships, modeled externally in a user experiment rather than extracted from training data. Results for initial concept detection are factorized based on semantic smoothness and adjusted in compliance with the extracted concept correlations. Refinement performance is demonstrated in experiments to show the effectiveness of our algorithm and the extracted correlations.

Keywords: Semantic indexing, concept refinement, detection refinement, semantic smoothness, lifelogging.

1 Introduction

Measuring human behavior can be carried out by utilizing sensors embedded in smartphones or wearable devices. This has enabled high-resolution insights into human behavior for quantified-self applications. Lifelogging [2] is the term used to describe the process of automatically, and ambiently, digitally recording human behavior using a variety of sensor types. The richness of contextual sensor information makes it practical to characterise and measure occurrences of human behavior [15] and in so doing to observe behavior patterns.

Visual media in digital format is now widespread and *visual lifelogging* which employs wearable cameras to capture image or video of everyday activities, has become popular in quantified-self applications. As a new form of multimedia content, the visual lifelog presents many challenges if its content is to be managed in a way that helps maximize benefit for the eventual applications. This requires searching through lifelogs based on content, and for this the automatic detection of semantic concepts is needed. The conventional approach to content-based indexing, as taken in the annual TRECVid benchmarking [11, 12], is to annotate a collection covering both positive and negative examples of the presence of each concept and then to train a machine learning classifier to recognize the presence of the concept. This typically requires a classifier for each concept without considering inter-concept relationships or dependencies yet in reality, many concept pairs will co-occur rather than occur independently.

In contrast to isolated concept detectors, *multi-label training* tries to classify and model correlations between concepts simultaneously. A typical multi-label training method is presented in [10], in which concept correlations are modeled in the classification model using Gibbs random fields. Similar multi-label training methods can be found in [20]. Since all concepts are learned from one integrated model, one shortcoming is the lack of flexibility, which means that the learning stage needs to be repeated when the concept lexicon is changed. Another disadvantage is the high complexity when modeling pairwise correlations in the learning stage. This also hampers the ability to scale up to large-scale sets of concepts and to complex concept inter-relationships.

As an alternative, *detection refinement or adjustment* methods post-process detection scores obtained from individual detectors, allowing independent and specialized classification techniques to be leveraged for each concept. Context-Based Concept Fusion (CBCF) is an approach to refining the detection results for independent concepts by modeling relationships between them [3]. Concept correlations are either learned from annotation sets [4] or in earlier work, inferred from pre-constructed knowledge bases [19, 5] such as WordNet. However, annotation sets are almost always inadequate for learning correlations due to their limited sizes and the annotation having being done with independent concepts rather than correlations in mind. The use of external knowledge networks also limits the flexibility of CBCF because they use a static lexicon which is costly to create. When concepts do not exist in an ontology, these methods cannot adapt to such situations.

To deal with such challenges, a training-free refinement method was proposed in [18] to post-process the outputs of individual concept detectors, independent of training corpora. While ontological relations can be employed in this method, the concept graph cannot be integrated for joint refinement in terms of semantic smoothness.

In this paper we propose a semantically smoothed refinement algorithm which extends our previous work and exploits concept relationships modeled externally in a user experiment rather than extracted from training data. As a training-free method, this algorithm can exploit inherent co-occurrence patterns for concepts

which exist in the detection results, exempt from the restrictions of correlation learning from a training corpus. Results for initial concept detection are factorized and adjusted in compliance with the concept correlations.

The contribution of our work is three-fold. First, we present an algorithm based on weighted non-negative matrix factorization to improve concept detection accuracy and we then semantically smooth the factorization procedure to constrain the results to comply with the concept correlations. Second, in contrast to training-dependent methods which learn concept correlations from training corpora, this paper tries to deal with concept refinement problem in a training-free way. Our concept correlations are inferred from a simple user experiment to exploit topic-related concept semantics. Finally, we compare our approach with various methods on different concept detection accuracy levels and we demonstrate our method to be advantageous in different ways.

2   Semantically Smoothed Everyday Concept Refinement

2.1   Motivation and Problem Statement

Our work is based on concept detection results from a series of images taken from lifelog events which have been automatically segmented based on the technique introduced in [8]. A lifelog event corresponds to a single activity in the wearer's day such as watching TV, commuting, or eating a meal, with an average stream of 20 or more events of varying duration in a typical day.

Let $\{E_1, E_2, ..., E_n\}$ be a set of event streams, event E_i is represented by successive images $I^{(i)} = \{Im_1^{(i)}, Im_2^{(i)}, ..., Im_k^{(i)}\}$. We assume a universe of concepts, C. Each image $Im_j^{(i)}$ might have several concepts detected, we assume the concepts appearing in image $Im_j^{(i)}$ are represented as a vector $C_j^{(i)} = \{c_{j1}^{(i)}, c_{j2}^{(i)}...c_{jM}^{(i)}\}$ for M concepts. The set of passively captured images in a lifelog is denoted as $I = \{I^{(1)}, I^{(1)}, ..., I^{(n)}\}$ which has dimensionality $N = \sum_{i=1}^{n} k_i$, 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 $C_{N \times M}$.

The semantic refinement task is to modify the $N \times M$ dimensional matrix C in order to keep consistency with the underlying contextual pattern of concepts. According to [7], matrix C can be represented as $C \approx WH$, in which W and H represent image-specific and concept-specific latent features respectively. The intuition behind this is to form the confidence matrix by simply combining partial information (columns in W) 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 encoding. If we constrain the factorization with these inferred concept-concept correlations to comply with the external concept graph, the factorized results can better reflect this contextual information which leads to improved refinement.

2.2   Refinement Specification

Let the dimensions of component matrix W and H be $N \times r$ and $r \times M$. The approximation factorization defined above can be solved by optimizing the cost function defined to qualify the quality of the approximation. In factorizing the confidence matrix, the weighted measure cost function is used. To distinguish the contribution of different concept detectors to the cost function, the weighted Frobenius norm [18, 9] can employed as:

$$G(W, H) = \frac{1}{2} \sum_{ij} w_{ij}(c_{ij} - W_{i.}H_{.j})^2 + \frac{\lambda}{2}(\|W\|_F^2 + \|H\|_F^2) \tag{1}$$

such that $W \geq 0, H \geq 0$ and where $\| \cdot \|_F^2$ denotes the Frobenius norm. The role of the quadratic regularization term $\lambda(\|W\|_F^2 + \|H\|_F^2)$ is to prevent overfitting.

Because each value c_{ij} in C denotes the probability of the occurrence of concept v_j in sample c_i, the estimation of the existence of v_j is more likely to be correct when c_{ij} is high, which is also adopted by [18] under the same assumption that the initial detectors are reasonably reliable if the returned confidences are larger than a threshold *thres*. To distinguish contributions of different concept detectors to the cost function, we employ a weighted matrix $W = (w_{ij})_{N \times M}$ whose elements are larger for reliable and lower for less reliable detections. After factorization, refinement can be expressed as a fusion of confidence matrices:

$$C' = \alpha C + (1 - \alpha)\tilde{C} = \alpha C + (1 - \alpha)WH \tag{2}$$

2.3   Semantic Smoothness

According to the factorization described in Section 2.2, matrix $H_{r \times M}$ represents a low rank latent feature matrix for M concepts. Therefore, concept correlations can be formalized based on this new form of concept representation. We also denote a smooth function S in weighted form to constrain two concepts with high correlation to have coordinates which are close to each other. Under this rationale, function S can be calculated as

$$\begin{aligned}
S(H) &= \frac{1}{2} \sum_{ij} \|H_{.i} - H_{.j}\|_F^2 Correl(c_i, c_j) \\
&= \sum_{ij} H_{.i}^T Correl(c_i, c_j)H_{.i} - \sum_{ij} H_{.i}^T Correl(c_i, c_j)H_{.j} \\
&= \sum_{ij} H_{.i}^T D_{ii}H_{.i} - \sum_{ij} H_{.i}^T Correl(c_i, c_j)H_{.j} \\
&= tr(H(D - Correl)H^T) = tr(HL_{Correl}H^T) \tag{3}
\end{aligned}$$

where $Correl$ is a concept correlation matrix whose element $Correl(c_i, c_j)$ stands for the quantitative relationship between two concepts c_i and c_j, $D_{ii} = \sum_j Correl(c_i, c_j)$ stands for a diagonal matrix, $L_{Correl} = D - Correl$ represents the Laplacian of

the concept correlation matrix. By assembling the above defined functions, we can achieve the following semantically smoothed factorization problem:

$$min_{W,H} F(W, H) = G(W, H) + \frac{\beta}{2} S(H) \qquad \text{s.t. } W, H \geq 0 \qquad (4)$$

where β is a nonnegative scalar to weight the penalty of concept correlations in the optimization of $F(W, H)$. In Eqn. (4), $G(W, H)$ interacts with $S(H)$ by sharing the concept feature matrix H, through which the factorization is optimized in compliance with concept correlation $Correl$. In this way, $Correl$ influences concept feature matrix H in terms of its Laplacian matrix L_{Correl}.

Because there is no close-form solution for optimizing Eqn. (4), we employed the classical gradient descent method to find a local optimization, implemented by updating W and H in the opposite direction to the gradient at each iteration. Note that the gradient of $S(H)$ with respect to H is $\nabla_H S(H) = 2HL_{Correl}$. Based on the above formalization, the partial derivations of F with respect to W and H can be calculated respectively by

$$\frac{\partial F}{\partial W} = [(WH - C \circ W]H^T + \lambda L \qquad (5)$$

$$\frac{\partial F}{\partial H} = W^T[(WH - C) \circ W] + \beta H L_{Correl} + \lambda H \qquad (6)$$

where $\circ$ denotes Hadamard multiplication. The semantically smoothed factorization solution of W and H is further applied to Eqn. (2) to calculate C' which can then be further enhanced by neighborhood-based propagation, shown in [18].

3   Inferring External Concept Correlations

In domains where context is very important, such as lifelogging, concepts are more contextually correlated within single event topics such as "talking", "cooking", etc. Concepts can play different roles in representing event semantics, and some will interact with each other through their relationships. This means that if we plot the concepts in a vector space, the dimensions in a concept vector are not independent because of their relationships to each other which are intermediated by event topic semantics. As shown in Fig. 1, the event semantic space can be defined as a linear space with a set of concepts as the bases. Since concepts are highly connected through event topics, we infer concept correlations through a topic-related method.

To investigate quantitative correlations among a set of candidate concepts related to everyday activities, we carried out user experiments on concept rating where candidate concepts related to each of the activity topics were pooled based on user input. Although individual wearers may have different contexts and personal characteristics, there is a common understanding of concepts that is already socially agreed and allows people to communicate about these according to [6] and [17]. This makes it reliable for users to choose suitable concepts relevant to activities. User experiments were carried out to discover candidate concepts which potentially have high correlation with activity semantics.
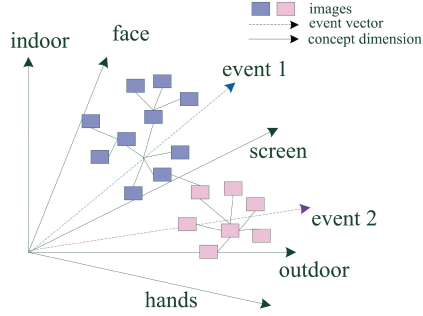
Fig. 1. Topic-related concept correlations in concept space.

Fig. 2. SenseCam worn by a participant

A group of 13 participants were chosen from among researchers within our own group, some of whom log their own everyday lives with the SenseCam wearable camera [13], shown in Fig. 2. In the experiment, target activities were first described to the participants to familiarise them. They were then shown SenseCam images for selected activity examples and surveyed by questionnaire about their interpretation of the SenseCam activity images as well as of the concepts occurring regularly in those SenseCam images. The aim of the user experiment was to determine candidate semantic concepts which have high correlation with various human activities. After several iterations and refinements we selected 85 base concepts [17, 15] which have highest agreement among participants, i.e. more than half of respondents think each are relevant to the activity.

Using rating results from the user experiment, concept correlation can be calculated based on the intuition that the higher the number of subjects who thought a concept pair was relevant to a given activity, the implication of a stronger concept correlation. By summation across all target activity categories, the aggregated correlation score can be quantified. The correlation score of a concept pair c_i and c_j is calculated by

$$Correl(c_i, c_j) = \frac{\sum_{act \in A} min(v_{act}(c_i), v_{act}(c_j))}{\sum_{act \in A} v_{act}(c_i) * \sum_{act \in A} v_{act}(c_i)}), i \neq j \qquad (7)$$

where A denotes the set of activities investigated in the user experiment. $v_{act}(c_i)$ is the number of votes concept c_i gets with regard to a specific activity type act, and $min(v_{act}(c_i), v_{act}(c_j))$ reflects the number of overlapping votes for a concept pair in activity type act. Since higher overlapping votes implies stronger agreement among all subjects on the relevance of two concepts within the context of a specific activity, the correlation of two concepts should be higher if the sum of overlapping votes are higher across all activity types. $\sum_{act \in A} v_{act}(c_i)$ and $\sum_{act \in A} v_{act}(c_j)$ denote the total vote numbers of c_i and c_j in all activities, which are used to normalize the correlation values.

4 Experiments and Evaluation

4.1 Experimental Dataset

To assess the performance of our algorithm, we used a set of 85 everyday concepts [17, 15] introduced in Section 3. We used a dataset including event samples of 23 activity types collected from 4 SenseCam wearers and consisting of 12,248 SenseCam images [15]. Concept detectors with different accuracy levels were used and mean average precision (MAP) calculated for evaluation. To test the performance on different levels of concept detection accuracy, detectors were simulated using the *Monte Carlo* method following the work in [1]. Various original concept detection accuracies were provided by varying the controlling parameter μ_1 in the range [0.5...10.0]. For each setting we executed 20 repeated runs and the averaged concept MAP was calculated. Fig. 3 shows the improvement in concept MAP with increasing μ_1. Near-perfect detection performances are achieved when $\mu_1 \geq 5.5$.
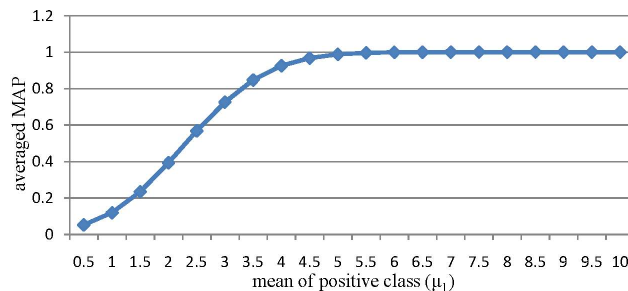

Fig. 3. Averaged concept MAP with different controlling parameter μ_1.

4.2 Evaluation and Discussion

In Table 1, our semantically smoothed refinement (SSR) method is compared with a variety of concept detection refinement methods including ontological refinement [19], a Random Walk-based method [14], Tensor-based refinement for wearable sensing [16], training-free refinement (TFR) [18], as well as domain adaptive semantic diffusion (DASD) [4]. We applied the same ontological structure of 85 concepts with *subsumption* and *disjointness* concept relationships as used in [18] and applied to ontological refinement. Motivated by [18], we implement the SSR in a training-free manner followed with neighbourhood-based propagation as introduced by [18] after semantically smoothed factorization. That is, applying a propagation algorithm to C' to further refine the results based on neighbors connected to each samples. In DASD, the concept correlation $Correl$ as generated in Section 3 is employed as the semantic graph used

in the diffusion. In this experiment, we employ the same parameter settings as reported in TFR method [18] during the factorization procedure. That is, we choose the number of latent features as $r = 10$ and we threshold the detection results with $thres = 0.3$. The fusion parameter in Eqn. (2) is simply set to $\alpha = 0.5$, assigning equal importance to the two matrices. We use 30 nearest neighbours in the propagation step.

Table 1. MAP of SSR, TFR, Ontological, Random Walk, Tensor and DASD (mean over 20 runs).

Method	$\mu_1 = 1.0$	$\mu_1 = 1.5$	$\mu_1 = 2.0$	$\mu_1 = 2.5$	$\mu_1 = 3.0$	$\mu_1 = 3.5$	$\mu_1 = 4.0$
Onto	0.159	0.273	0.421	0.586	0.735	0.850	0.926
RW	0.156	0.267	0.426	0.603	0.752	0.857	0.924
Tens	0.164	0.287	0.456	0.624	0.774	0.877	0.941
TFR	0.165	0.295	0.477	0.658	0.800	**0.893**	**0.947**
DASD	**0.176**	0.300	0.454	0.602	0.719	0.805	0.869
SSR	0.168	**0.302**	**0.480**	**0.662**	**0.802**	**0.893**	0.944

From Table 1, we find that SSR outperforms the other methods at most levels of original detection MAP from 0.15@μ_1 = 1.0 to 0.92@μ_1 = 4.0. At $\mu_1 = 1.0$, SSR is outperformed by DASD. This makes sense as initial detection accuracy is low. In this case, very few correctly detected concepts are selected by thresholding for further refinement which is impractical in real world applications. Though DASD performs well at $\mu_1 = 1.0$, it is significantly outperformed by TFR and SSR methods at $\mu_1 \geq 2.0$. This is probably because the concept correlation graph is inferred from the user experiment rather than from the training dataset. Thus the knowledge-driven correlation graph only reflects the co-occurrence tendency which is difficult to precisely incorporate in DASD refinement. When the original concept detection accuracy is low such as at $\mu_1 = 1.0$, the overfitting of DASD to semantic graph $Correl$ can enhance the results but this can not be generalized to other concept detection accuracy levels. However, SSR can adapt to various levels of concept detections and make full use of external knowledge graph $Correl$. Compared to the TFR method, SSR incorporates the imprecisely constructed concept graph but outperforms TFR in most cases. This demonstrates the potential of SSR in everyday concept refinement through semantic smoothness. Recall that we employed a flexible but imprecisely constructed concept correlation graph for smoothness. We believe that if the correlation can be inferred more precisely from data annotations, the improvement will be more significant with the semantic smoothness.

To further demonstrate the effects of concept correlations on indexing refinement, we compared the semantically smoothed refinement results to non-smoothed results ($\beta = 0$) by varying parameter β which controls the influence of concept correlations. As we can see from Eqn. (4), the SSR method degenerates to TFR at $\beta = 0$. In Fig. 4 and 5, comparisons at detection accuracy controlled by $\mu_1 = 2.0$ are shown for refinement results without and with neighborhood
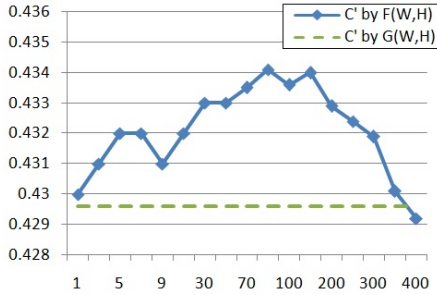
Fig. 4. Effects of concept correlations without propagation ($@\mu_1 = 2.0$).
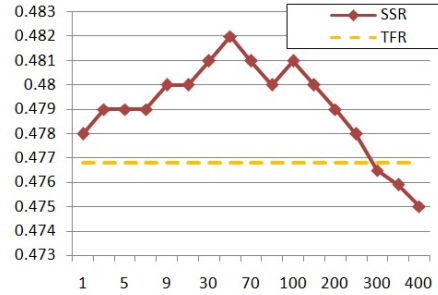
Fig. 5. Effects of concept correlations with propagation ($@\mu_1 = 2.0$).

propagation respectively. In Fig. 4, the blue dashed line represents the refinement results evaluated on the intermediate results of C' calculated by optimizing non-smoothed function $G(W, H)$. As we can see from Fig. 4, the results obtained from smoothed function $F(W, H)$ surpasses $G(W, H)$ in a wide range of β. While β increases, the performance using a smoothed function $F(W, H)$ climbs first, and then declines when the contribution of concept correlations dominate the factorization with very large β. This can interpret the overfitting of the DASD method in which the cost function is defined on the concept graph [4] for semantic diffusion, similar to the function $G(W, H)$ we used. When further applying neighborhood-based propagation to C', both smoothed and non-smoothed methods can be boosted, as shown in Fig. 5 in which the final enhanced results of SSR and TFR are compared. As in Fig. 4, SSR still surpasses TFR in most cases in Fig. 5 showing the advantages of semantic smoothness even if neighborhood-based propagation is applied.

5   Conclusions and Future Work

We tackle the multi-concept refinement task in quantified-self like visual lifelogging where activity engagements of users usually dominate the contextual evolution of concept appearance. We propose a semantically smoothed refinement algorithm to exploit topic-related concept relationships to further improve semantic indexing performance. Instead of extracting concept correlations from annotated training data, which is usually employed in other research but might be constrained by the labeling quality of a training corpus, we investigated a simple method of externally generating concept graphs in a user experiment. Though loosely constructed, the concept correlation graph has been demonstrated as effective when incorporated into the smoothed refinement algorithm. Incorporating automatic generation of concept graphs from a knowledge base or other sources is future work to extend the algorithm.

References

1. Aly, R., Hiemstra, D., de Jong, F., Apers, P.: Simulating the future of concept-based video retrieval under improved detector performance. Multimedia Tools and Applications pp. 1–29 (2011)
2. Gurrin, C., Smeaton, A., Doherty, A.: LifeLogging: Personal Big Data. Foundations and Trends in Information Retrieval 8(1), 1–127 (2014)
3. Jiang, W., Chang, S.F., Loui, A.: Context-based concept fusion with boosted conditional random fields. In: ICASSP 2007. vol. 1, pp. I–949–I–952 (April 2007)
4. Jiang, Y.G., Dai, Q., Wang, J., Ngo, C.W., Xue, X., Chang, S.F.: Fast semantic diffusion for large-scale context-based image and video annotation. IEEE Transactions on Image Processing 21(6), 3080–3091 (June 2012)
5. Jin, Y., Khan, L., Wang, L., Awad, M.: Image annotations by combining multiple evidence & WordNet. In: ACM MULTIMEDIA'05. pp. 706–715 (2005)
6. Lakoff, G.: Women, Fire, and Dangerous Things. Univ. Chicago Press (April 1990)
7. Lee, D., Seung, H.: Learning the parts of objects by nonnegative matrix factorization. Nature 401, 788–791 (1999)
8. Lee, H., Smeaton, A., O'Connor, N., Jones, G., Blighe, M., Byrne, D., Doherty, A., Gurrin, C.: Constructing a SenseCam visual diary as a media process. Multimedia Syst. 14(6), 341–349 (2008)
9. Pan, R., Zhou, Y., Cao, B., Liu, N.N., Lukose, R., Scholz, M., Yang, Q.: One-class collaborative filtering. In: ICDM'08. pp. 502–511. IEEE (2008)
10. Qi, G.J., Hua, X.S., Rui, Y., Tang, J., Mei, T., Zhang, H.J.: Correlative multi-label video annotation. In: ACM MULTIMEDIA'07. pp. 17–26 (2007)
11. Smeaton, A.F., Over, P., Kraaij, W.: Evaluation campaigns and trecvid. In: Proceedings of the ACM international workshop on Multimedia Information Retrieval. pp. 321–330. ACM (2006)
12. Smeaton, A.F., Over, P., Kraaij, W.: High level feature detection from video in TRECVid: a 5-year retrospective of achievements. In: Ajay Divakaran (Ed.), Multimedia Content Analysis, Theory and Applications. pp. 151–174 (2008)
13. Steve, H., Emma, B., Ken, W.: Sensecam: a wearable camera that stimulates and rehabilitates autobiographical memory. Memory 19(7), 685–96 (2011)
14. Wang, C., Jing, F., Zhang, L., Zhang, H.J.: Image annotation refinement using random walk with restarts. In: ACM MULTIMEDIA'06. pp. 647–650 (2006)
15. Wang, P., Smeaton, A.: Using visual lifelogs to automatically characterise everyday activities. Information Science 230, 147–161 (2013)
16. Wang, P., Smeaton, A., Gurrin, C.: Factorizing time-aware multi-way tensors for enhancing semantic wearable sensing. In: 21st International Conference, MMM 2015. pp. 571–582 (Jan 2015)
17. Wang, P., Smeaton, A.F.: Semantics-based selection of everyday concepts in visual lifelogging. Intnl. Journal of Multimedia Information Retrieval 1(2), 87–101 (2012)
18. Wang, P., Sun, L., Yang, S., Smeaton, A.F.: MultiMedia Modeling: 22nd International Conference, MMM 2016, Miami, FL, USA, January 4-6, 2016, Proceedings, Part I, chap. Towards Training-Free Refinement for Semantic Indexing of Visual Media, pp. 251–263. Springer International Publishing, Cham (2016)
19. Wu, Y., Tseng, B., Smith, J.: Ontology-based multi-classification learning for video concept detection. In: ICME'04. vol. 2, pp. 1003–1006 Vol.2 (2004)
20. Xue, X., Zhang, W., Zhang, J., Wu, B., Fan, J., Lu, Y.: Correlative multi-label multi-instance image annotation. In: ICCV 2011. pp. 651–658 (Nov 2011)