Training-Free Indexing Refinement for Visual Media via Multi-Semantics

Peng Wang [1]*, Lifeng Sun[1]*, Shiqiang Yang[1]*, Alan F. Smeaton[2]*

[1] *National Laboratory for Information Science and Technology*
Department of Computer Science and Technology
Tsinghua University, Beijing, 100084, China

[2] *Insight Centre for Data Analytics*
Dublin City University, Glasnevin, Dublin 9, Ireland

Abstract

Indexing of visual media based on content analysis has now moved beyond using individual concept detectors and there is now a focus on combining concepts by post-processing the outputs of individual concept detection. Due to the limitations and availability of training corpora which are usually sparsely and imprecisely labeled with concept groundtruth, training-based refinement methods for semantic indexing of visual media suffer in correctly capturing relationships between concepts, including co-occurrence and ontological relationships. In contrast to training-dependent methods which dominate this field, this paper presents a training-free refinement (TFR) algorithm for enhancing semantic indexing of visual media based purely on concept detection results, making the refinement of initial concept detections based on semantic enhancement, practical and flexible. This is achieved using what can be called multi-semantics, factoring in semantics from multiple sources. In the case of this paper, global and temporal neighbourhood information inferred from the original concept detections in terms of weighted non-negative matrix

*Corresponding author at: Department of Computer Science and Technology, Tsinghua University, Beijing, 100084, China. Tel: +86 -10 -62786910

Email addresses: pwang@tsinghua.edu.cn (Peng Wang [1]), sunlf@tsinghua.edu.cn (Lifeng Sun[1]), yangshq@tsinghua.edu.cn (Shiqiang Yang[1]), alan.smeaton@dcu.ie (Alan F. Smeaton[2])

factorization and neighbourhood-based graph propagation are both used in the refinement of semantics. Furthermore, any available ontological concept relationships among concepts can also be integrated into this model as an additional source of external *a priori* knowledge. Extended experiments on two heterogeneous datasets, images from wearable cameras and videos from TRECVid, demonstrate the efficacy of the proposed TFR solution.

Keywords: Semantic indexing, Refinement, Concept detection enhancement, Context fusion, Factorization, Propagation

1. Introduction

Video in digital format is now in widespread use in everyday scenarios. While mainstream consumer-based access to image and video on platforms such as YouTube and Vine are based on user tags and metadata, prevailing methods to indexing based on *content* detect the presence or absence of semantic concepts which might be general (e.g., *indoor*, *face*) or more abstract (e.g., *violence*, *meeting*). The conventional approach to content-based indexing of visual media, as taken in the annual TRECVid benchmarking [21, 20], is to manually annotate a collection of visual media covering both positive and negative examples, for the presence of each concept. This can be done manually, or can use visual captchas [16], and then train a machine learning classifier using these annotations to recognise the presence, or absence, of the semantic concept. This typically requires a classifier for each concept without considering inter-concept relationships or dependencies yet in reality, many concept pairs and triples are often semantically related and dependent and thus will co-occur rather than occur independently. It is widely accepted and it is intuitive that detection accuracy for concepts can be improved if concept correlation can be exploited.

The idea of refining an initial, raw, set of concept detections is intuitive and has been explored for some time and it is still currently a topic attracting a lot of attention, such as in [14]. Context-Based Concept Fusion (CBCF) is an approach to refining the detection results for independent concepts by modeling relationships between them [5]. Concept correlations are either learned from annotation sets [10, 24, 25, 8, 6] or inferred from pre-constructed knowledge bases [28, 9] 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. In addition, training sets may not be fully labeled or may be noisy. The use of external knowledge networks also limits the flexibility of CBCF because it uses a static lexicon which is costly to create and even costlier to maintain. When concepts do not exist in an ontology, these methods cannot adapt to such situations.

In this paper we propose a training-free refinement (TFR) method to exploit inherent co-occurrence patterns for concepts which exist in testing sets, exempt from the restrictions of training corpus and external knowledge structures and we use this to refine and improve the output of independent concept classifiers. TFR can fully exploit various sources of semantic information including global patterns of multi-concept appearance, an ontology encapsulating any concept relations (if available), as well as sampling the distribution of concept occurrences in the temporal neighbourhood of a given image, all with the goal to enhance the original one-per-class concept detectors and all done within a unified framework. Although this reduces the learning/training process, we set out here to see if TFR can still obtain better or comparable performance than the state-of-the-art as such an investigation into refinement of semantic indexing has not been done before.

The contributions of this paper can be highlighted as:

- A training-free refinement method which uses information inferred from test datasets without any requirement for high quality training data based on full concept annotations. This can flexibly adapt to many real world applications where only limited or incomplete annotations are available for correlation inference and goes beyond the state-of-the-art in that it is flexible and dynamically adaptable to new domains or datasets, without the need for a training phase;

- An ontological factorization algorithm to adjust and improve on the initial less accurate results for concept detection, according to the global patterns of concept appearance and absence, across the whole collection of samples. Ontology-based concept relationships can also be combined into this algorithm as another source of external *a priori* knowledge thus illustrating how the TFR method presented here, can easily incorporate new sources of evidence for concept refinement, unlike other available approaches;

- A similarity graph of nearest neighbours based on the refined results using ontological factorization and applying a graph propagation algorithm to further enhance the detection accuracy exploiting such local relationships, which finally achieves satisfactory refinement, something which has not been available previously;

- A set of experiments on two heterogeneous datasets, chosen to validate the effectiveness of the above.

The rest of the paper is organized as follows: in Section 2 we review related work on refinement of semantic indexing. In Section 3 we present an overview of our TFR solution and algorithm followed by a detailed elaboration of TFR in Section 4. A set of experiments including a description of the two datasets we used and a discussion of results, are presented in Section 5. We finish with conclusions and proposals for future work.

2. Related Work

The task of automatically determining the presence or absence of a semantic concept in an image or a video shot (or a keyframe) has been the subject of at least a decade of intensive research. The earliest approaches treated the detection of each semantic concept as a process independent of the detection of other concepts and used supervised learning approaches to implement this, but it was quickly realised that such an approach is not scalable to large numbers of concepts, and does not take advantage of inter-concept relationships. Based on this realisation, there have been efforts within the multimedia retrieval community focusing on utilization of inter-concept relationships to enhance detection performances, which can be categorized into two paradigms: multi-label training and detection refinement or adjustment.

In contrast to isolated concept detectors, *multi-label training* tries to classify concepts and to model correlations between them, simultaneously. A typical multi-label training method is presented in [18], in which concept correlations are modeled in the classification model using Gibbs random fields. Similar multi-label training methods can be found in [30]. 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.

There has also been some work on *multi-label detection*, within the framework of TRECVid where for the 2012 and 2013 edition of the TRECVid semantic indexing task, a secondary "concept pair" task was offered. The motivation here is a video (but could equally well be image) retrieval scenario which demands complex queries that go beyond a single concept. Examples of concept pairs which could go together include $Animal + Snow$, $Person + Underwater$ and $Boat/Ship + Bridges$. Rather than combining concept detectors at query time, the TRECVid concept pair task aimed at detecting the simultaneous occurrence of a pair of unrelated concepts in a video.

In 2012 the top run achieved a score of 0.076 MAP and in 2013 the top run achieved a score of 0.162 MAP [2]. While this seems an improvement, it should be noted that the pairs changed from one year to the next and some may have been easier, or less rare, than the ones in 2012. Of course there was variability in performance across concept pairs but the best performer for the pair $Government\ Leader + Flags$, for example, scored 0.658 MAP

which is very respectable.

The approaches taken by various participants in this activity were mostly based around combining multiple individual detectors by well known fusion schemes, including sum, product and geometric mean and while it represents an interesting exploration, the feasibility of indexing visual media, at indexing time, by concept pairs and scaling this to large collections would seem remote.

As an alternative to concept detection at indexing time, *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. Detection refinement has attracted interest based on exploiting concept correlations inferred from annotation sets [10, 24, 25, 5] or from pre-constructed knowledge bases [28, 9, 12]. However, these depend on training data or external knowledge. When concepts do not exist in the lexicon ontology or when extra annotation sets are insufficient for correlation learning as a result of the limited size of the corpus or of sparse annotations, these methods cannot adapt to such situations. Another difficulty is the matter of determining how to quantify the adjustment when applying the correlation. Though concept similarity [9], sigmoid function [28], mutual information [10], random walk [24, 25], random field [5], etc. have all been explored, this is still a challenge in the refinement of concept detections. In a state-of-the-art refinement method for indexing TV news video [8, 6], the concept graph is learned from the training set. Though adaptation is considered to handle changes between training and test data, the migration of concept alinement to testing sets also depends on the affinity of two data sets, which is not always the case and can reduce the performance of indexing user-generated media, for example. Moreover, incomplete or imprecise annotations on training sets will further degrade the performance of these methods which rely highly on inter-concept correlations learned from training labels. The proposed TRF method in this paper is indeed a refinement methods but tries to tackle the above challenges.

These approaches to improving concept detection all try to compensate for the fact that it is really difficult to get accurate training data, i.e. annotations. TRECVid, the largest collaborative benchmarking activity in the area, with its collaborative annotation of training data among participants in one year realised a total of 8,158,517 annotations made directly by the participants of TRECVid or by the annotators of the Quaero project and a total of 28,864,844 annotations was obtained by propagating the initial annotations using the *implies* or *excludes* relations among concepts. While this may appear substantial and used clever techniques like an active learning procedure to prioritise annotations of the most useful sample shots [3] and to ask for a "second opinion" when manual annotations strongly disagreed with a prediction [19], this was still for only 346 concepts in TRECVid 2010 to 2015. Clearly this is not sustainable to a larger and more realistic set of concepts so between 2012 and 2015 a "no annotation" task was offered in TRECVid, to reflect the difficulty associated with finding good training data for the supervised learning tools which have become commonplace.

The potential for automatically harvesting annotations or training data for supervised learning from web resources has been recognised by many, including the first such work by [23]. While participation in this aspect of the semantic indexing task in TRECVid was low, by 2014 the best submission scored 0.078 in terms of MAP against a best submission using manual annotations of 0.34 MAP, quite a long way behind [2]. While these results are encouraging, much more work remains to be done in this area.

3. Motivation and Proposed Solution

Fusing the results of concept detection to provide better quality semantic analysis and indexing is a challenge. Current research is focused on learning inter-concept relationships explicitly from training corpora and then applying these to test sets. Since the initial results of semantic concept detection will always be noisy because of the accuracy level at which they operate, little work has investigated a refinement approach which directly uses the original detection results to exploit correlations. However, according to the TRECVid benchmark, acceptable detection results can now be achieved, particularly for concepts for which there exists enough annotated training data [20, 22, 2]. These detections with high accuracies should be used as cues to enhance overall multi-concept detections since the concepts are highly correlated, though the bottleneck is in the correlation itself which is difficult to precisely model.

For much of the visual media we use in our everyday lives there is a temporal aspect. For example video is inherently temporal as it captures imagery over time and thus video shots or keyframes from shots may have related content because they are taken from the same scene or have the same characters of related activities. Likewise still images of a social event captured in sequence will have semantic relationships based on shared locations, activities or people. We represent these related samples in terms of "neighbors" which are likely to be similar within the same time range. For such "connected" visual media it makes sense to try to exploit the temporal relationships when post-processing initial concept detection, and to use the "neighbourhood" aspect of visual media.

Our TFR method is thus motivated based on the following:

- **Reliability**: Detection results for at least some concepts should be accurate enough to be exploited as reliable cues for a refinement process.
- **Correlation**: Instead of occurring in isolation, concepts usually co-occur or occur mutually exclusively among the same samples.
- **Compactness**: Since concept occurrences are not fully independent, detection results can be projected to a compact semantic space.
- **Re-Occurrence**: Concepts will frequently occur across semantically similar samples so where the visual media has temporal relationships such as video keyframes, neighbourhood relationships can be exploited.

Based on the above motivations, the TFR method is proposed which will combine the correlation of individual concepts with various detection accuracies, to improve the performance of overall semantic indexing. The overview of this proposed solution is illustrated in Fig. 1. In Fig. 1(a), initial concept detection is first applied to a set of visual media inputs, returning results denoted as matrix C where each row $s_i (1 \le i \le N)$ represents a sample media element such as an image or video shot, while each column corresponds to a concept $v_j (1 \le j \le M)$ in the vocabulary. We use different gray levels to represent matrix elements in C, namely the confidences of concept detections.

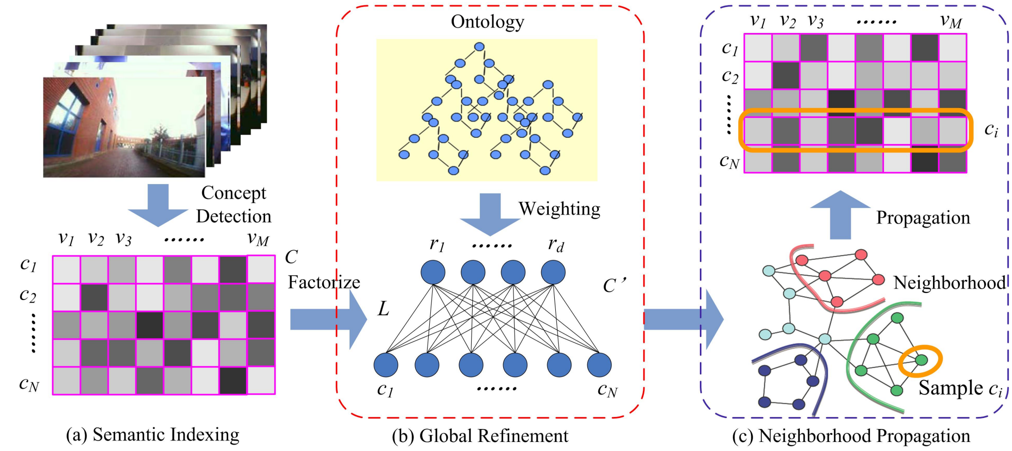

Figure 1: Illustration of the TFR framework. (a) Semantic Indexing: Media samples indexed through concept detections, returning C. (b) Global Refinement (GR): Refining C as C' using global contextual patterns. (c) Neighbourhood propagation (NP): Refining C' by similarity propagation between nearest neighbours.

As shown in Fig. 1, the refinement procedure involves two stages of global refinement (GR) and neighborhood propagation (NP). The intuition behind GR is that, the high-probable correct detection results are selected to construct an incomplete but more reliable matrix which is then completed by a factorization method. Matrix factorization is one approach which has been used as a way to refine initial, usually automated, assignments of content descriptions or tags in work applied to social tags [13] or visual bag-of-words [14]. In our work, GR in Fig. 1(b) is a weighted matrix factorization process and performs an estimation of concept detection results which were less accurate in the original matrix C. If ontological relationships among concepts exist, they may also be employed to appropriately choose the entry value in the weighted matrix in correspondence to C. In Fig. 1(c), reconstructed concept detection results C' are used to calculate the sample-wise similarity in order to identify a number of nearest neighbours of the target sample s_i.

The propagation algorithm is then applied to infer labels iteratively based on neighbours connected to each sample.

4. Training-Free Refinement (TFR)

As illustrated in Fig. 1, GR and NP in the TFR framework are implemented by ontological factorization and graph propagation, which exploit global patterns and local similarities respectively.

4.1. Factorizing Detection Results

In GR, the task of detection factorization is to modify the $N \times M$ matrix C to overlay a consistency on the underlying contextual pattern of concept occurrences. Non-negative matrix factorization (NMF) has shown advantages in scalably detecting the essential features of input data with sparsity, which is more suitable to the semantic indexing refinement task where the annotations are sparse and the confidences in C are non-negative.

As distinct to the traditional NMF method, we need to optimize the factorization problem in weighted low ranks to reflect different accuracies of concept detections in GF. For this purpose, we employ a weight matrix $W = (w_{ij})_{N\times M}$ whose elements are larger for reliable, and lower for less reliable detections, to distinguish contributions of different concept detectors to the cost function. Because each value c_{ij} in C denotes the probability of the occurrence of concept v_j in sample s_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 [10, 26] under the same assumption that the initial detectors are reasonably reliable if the returned confidences are larger than a threshold. While we can simply assign $w_{ij} = 1$ for $c_{ij} \geq \mathit{threshold}$ and $w_{ij} \in (0, 1)$ uniformly for $c_{ij} < threshold$, we will describe a more sophisticated weighting scheme using ontologies in Section 4.2.

The application of weighted NMF here is to represent C as $\tilde{C} = LR$, where vectors in $L_{N\times d}$ and $R_{d\times M}$ can be referred to as d-dimensional sample-related and concept-related latent factors. By applying rules of customized optimization, each confidence value in C can be refined as $\tilde{c}_{ij} = \sum_{k=1}^{d} l_{ik} r_{kj}$. We define the following cost function and solve for L and R by optimizing the weighted least square form:

$$F = \frac{1}{2}\sum_{ij} w_{ij}(c_{ij} - L_{i\cdot}R_{\cdot j})^2 + \frac{\lambda}{2}(\|L\|_F^2 + \|R\|_F^2) \tag{1}$$

such that $L \geq 0, R \geq 0$ where $\|\cdot\|_F^2$ denotes the Frobenius norm and the quadratic regularization term $\lambda(\|L\|_F^2 + \|R\|_F^2)$ is applied to prevent over-fitting. After factorization, refinement can be expressed as a fusion of confidence matrices:

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

To solve the factorization problem, we use a multiplicative method [11] which has the advantage of re-scaling the learning rate instead of optimization with a fixed and sufficient small rate. Without loss of generality, we focus on the update of R in the following derivation and the update rule for L can be obtained in a similar manner. Inspired by [11], we construct an auxiliary function $G(r, r^k)$ of $F(r)$ for fixed L and each corresponding column r, c, w in R, C and W respectively. $G(r, r^k)$ should satisfy the conditions $G(r, r^k) \geq F(r)$ and $G(r, r) = F(r)$. Therefore, $F(r)$ is non-increasing under the update rule [11]:

$$r^{t+1} = argmin_r G(r, r^t) \tag{3}$$

where r^t and r^{t+1} stand for r values in two successive iterations. For function F defined in Eqn. (1), we construct G as

$$G(r, r^t) = F(r^t) + (r - r^t)^T \nabla F(r^t) + \frac{1}{2}(r - r^t)^T K(r^t)(r - r^t) \tag{4}$$

where r^t is the current update of optimization for Eqn. (1). Denoting $D(\cdot)$ as a diagonal matrix with elements from a vector on the diagonal, $K(r^t)$ in Eqn. (4) is defined as

$$K(r^t) = D(\frac{(L^T D_w L + \lambda I) r^k}{r^k}) \tag{5}$$

where $D_w = D(w)$ and the division is performed in an element-wise manner.

According to Eqn. (3), r can be updated by optimizing $G(r, r^t)$. By solving $\frac{\partial G(r,r^t)}{\partial r} = 0$, we obtain

$$\nabla F(r^t) + K(r^t) r - K(r^t) r^t = 0 \tag{6}$$

where

$$\nabla F(r^t) = L^T D_w (L r^t - c) + \lambda r^t \tag{7}$$

The combination of Eqn. (6) and (7) achieves the update rule

$$R_{kj}^{t+1} \leftarrow R_{kj}^t \frac{[L^T (C \circ W)]_{kj}}{[L^T (LR \circ W)]_{kj} + \lambda R_{kj}} \tag{8}$$

Similarly, each elements in matrix L can be updated by

$$L_{ik}^{t+1} \leftarrow L_{ik}^{t} \frac{[(C \circ W)R^T]_{ik}}{[(LR \circ W)R^T]_{ik} + \lambda L_{ik}} \tag{9}$$

where $\circ$ denotes Hadamard (element-wise) multiplication and each element in L can be updated similarly. According to Eqn. (3), the proof of $F(r)$ being non-increasing under the update rule given by Eqn. (8) and (9) is indeed the proof of $G(r, r^t)$ being an auxiliary function of $F(r)$, which is to be described in the analysis of the effectiveness of the approximation in Section 4.3.

4.2. Integration with Ontologies

In Section 4.1, we applied weighted NMF (WNMF) to perform low-accuracy concept estimation based on the assumption that the credibility of concepts in C is high enough if their detection confidence is larger than a predefined threshold. If we assign uniform weights for low-confidence concepts, WNMF will adjust confidences in terms of equal chance over these concepts. However, this is not the case in real world applications, where we often have biased estimations. To reflect concept semantics in W we introduce an ontological weighting scheme for WNMF-based global refinement.

To model concept semantics, an ontology is employed to choose appropriate weights for different concepts based on their semantics, similar in principle to the work reported in [31]. The goal is to correctly construct the matrix W which can reflect the interaction between concepts and their detection accuracy. Based on this motivation, we denote the ascendant concepts and descendant concepts for concept v as $ASC(v)$ and $DES(v)$. Similarly, the disjoint concepts explicitly modeled in the ontology are $DIS(v)$. The confidence of sample s belonging to concept v being returned by a detector is represented as $Conf(v|s)$. We introduce the multi-class margin factor [12] as

$$Conf(v|s) - max_{v_i \in D} Conf(v_i|s) \tag{10}$$

where D is the universal set of disjoint concepts of v which contains all concepts exclusively occurring with v. Note that $D \supseteq DIS(v)$ because there are also concepts modeled implicitly as disjoint with v in the ontology. For example, we only state "indoor" and "outdoor" are two disjoint concepts in an ontology and "tree", "sky" and "road" as descendant concepts of "outdoor". Then DIS(indoor) includes "outdoor" only, but all disjoint concepts of "indoor" include "outdoor" and all descendants of "outdoor" like "tree", "sky" and "road". Indeed, D includes $DIS(v)$ as well as $DES(DIS(v))$, which are all descendants of disjoint concepts of v, and disjoint concepts of ascendent concepts above v, denoted as $DIS(ASC(v))$. These statements of disjointness can be asserted or inferred. The former is created directly by the ontology to assert the statement. However, for the latter, a semantic reasoner is required to infer additional disjointness statements logically. Various reasoners such as RDFS [4] inference or OWL [15] inference can be embedded straightforwardly in our algorithm to leverage explicit statements to create logically valid but implicit statements.

By employing an ontology we assign each element in W as

$$w_{ij} \propto 1 - [c_{ij} - max_{v_k \in D} c_{ik}] \tag{11}$$

The interpretation of the weighting scheme is that if the disjoint concepts of v_j have higher detection confidences, it is less likely that v_j exists in sample s_i. In this case, the weight for concept v_j needs to be larger, otherwise the weight is lowered by ontology relationships using the multi-class margin.

4.3. Proof of Convergence

According to Eqn. (4), $G(r, r) = F(r)$ is satisfied and the proof of function $G(r, r^t)$ being an auxiliary of $F(r)$ is indeed the proof of $G(r, r^t) \geq F(r)$. For this purpose, we expand function $F(r)$ in the form of

$$\begin{aligned} F(r) &= \frac{1}{2}(c - Lr)^T D_w (c - Lr) + \frac{\lambda}{2} r^T r + C(L) \\ &= F(r^t) + (r - r^t)^T \nabla F(r^t) \\ &+ \frac{1}{2}(r - r^t)^T (L^T D_w L + \lambda I)(r - r^t) \end{aligned} \tag{12}$$

where I is $d \times d$ identity matrix and $C(L)$ is only relevant to L. According to Eqn. (4) and (12), we need to prove

$$(r - r^t)^T (K(r^t) - L^T D_w L - \lambda I)(r - r^t) \geq 0 \tag{13}$$

Substituting Eqn. (5) into (13), this is equal to proving that $D(\frac{L^T D_w L r^t}{r^t}) - L^T D_w L$ is positive semi-definite. We define a rescaling matrix as

$$\begin{aligned} M &= D(r^t)(D(\frac{L^T D_w L r^t}{r^t}) - L^T D_w L) D(r^t) \\ &= D(L^T D_w L r^t) D(r^t) - D(r^t)(L^T D_w L) D(r^t) \end{aligned} \tag{14}$$

For any vector v, since M is a symmetric matrix, we have

$$\begin{aligned} v^T M v &= \sum_{ij} v_i M_{ij} v_j \\ &= \sum_{ij} [r_i^t (L^T D_w L)_{ij} r_j^t v_i^2 - v_i r_i^t (L^T D_w L)_{ij} r_j^t v_j] \\ &= \sum_{ij} (L^T D_w L)_{ij} r_i^t r_j^t [\frac{1}{2} v_i^2 + \frac{1}{2} v_j^2 - v_i v_j] \\ &= \frac{1}{2} \sum_{ij} (L^T D_w L)_{ij} r_i^t r_j^t (v_i - v_j)^2 \geq 0 \end{aligned} \tag{15}$$

So far, we can conclude that $D(\frac{L^T D_w L r^t}{r^t}) - L^T D_w L$ is positive semi-definite, hence $G(r, r^t)$ is an auxiliary of $F(r)$. This guarantees effectiveness using the iterative update rules given in Eqn. (8) and (9).

4.4. Temporal Neighbourhood-Based Propagation

As shown in Fig. 1(c), temporal neighbourhood-based propagation further refines C' to achieve better indexing by exploiting local information between samples which are semantically similar. This procedure consists of two steps namely similarity-based neighbour localization and graph propagation.

4.4.1. Similarity Calculation

Following GR, detection results will have been adjusted in a way consistent with the latent sample/concept factors modeled in WNMF. While this procedure exploits general contextual patterns which are modeled globally by matrix factorization, the similarity propagation method can further refine the result by exploiting any local relationships between samples as demonstrated in Fig. 1(c). In this, it is important to localize highly related temporal neighbours for similarity-based propagation, for which the results C' after GR can provide better measures.

To derive the similarity between samples s_i and s_j, we calculate based on the refined results C' formulated in Eqn. (2) by Pearson Correlation, defined as:

$$P_{i,j} = \frac{\sum_{k=1}^{M}(c'_{ik} - \bar{c'}_i)(c'_{jk} - \bar{c'}_j)}{\sqrt{\sum_{k=1}^{M}(c'_{ik} - \bar{c'}_i)^2}\sqrt{\sum_{k=1}^{M}(c'_{jk} - \bar{c'}_j)^2}}$$

where $c'_i = (c'_{ik})_{1\leq k\leq M}$ is the i-th row of C', and $\bar{c'}_i$ is the average weight for c'_i. To normalize the similarity, we employ the Gaussian formula and denote the similarity as:

$$P'_{i,j} = e^{-\frac{(1-P_{i,j})^2}{2\delta^2}} \tag{16}$$

where δ is a scaling parameter for sample-wise distance. Based on this we can localize the k nearest neighbours of any target sample c_i which is highlighted with an orange circle in Fig. 1(c). Neighbours of c_i are indicated with green dots connected with edges quantified by Eqn. (16).

4.4.2. Graph Propagation

For implementing graph propagation, the NP procedure localizes k nearest neighbours for further propagation which are connected with the target sample in an undirected graph. The label propagation algorithm [29] is derived to predict more accurate concept detection results based on this fully connected graph whose edge weights are calculated by the similarity metric in Eqn. (16). Mathematically, this graph can be represented with a sample-wise similarity matrix as $G = (P'_{i,j})_{(k+1)\times(k+1)}$, where the first k rows and columns stand for the k nearest neighbours of a target sample to be refined which is denoted as the last row and column in the matrix. The propagation probability matrix T is then constructed by normalizing G at each column as

$$t_{i,j} = \frac{P'_{i,j}}{\sum_{l=1}^{k+1} P'_{l,j}}$$

which guarantees the probability interpretation at columns of T. By denoting the row index of k nearest neighbours of a sample c'_i to be refined as $n_i(1 \leq i \leq k)$ in C' and stacking the corresponding rows one below another, the neighbourhood confidence matrix can be constructed as $C_n = (c'_{n_1}; c'_{n_2}; ...; c'_{n_k}; c'_i)$. The propagation algorithm is carried out iteratively by updating

$$C_n^t \leftarrow \beta T C_n^{t-1} \tag{17}$$

where the first k rows in C_n stand for the k neighbourhood samples in C' indexed by subscript n_i and the last row corresponds to the confidence vector of the target sample c'_i. Since C_n is a subset of C', the graph G constructed on C_n is indeed a subgraph of the global graph constructed on C' as shown in Fig. 1(c). During each iteration, the neighbourhood concept vector c'_{n_i} needs to be clamped to avoid fading away. After a number of iterations, the algorithm converges to a solution in which the last row of C_n is a prediction based on similarity propagation. In this way, the local relationships between neighbours can be used for a more comprehensive refinement.

5. Experiments and Discussion

We assessed the performance of the TFR approach on two heterogenous datasets, a dataset of still images collected from wearable cameras (Dataset1) and the videos used in the TRECVid 2006 evaluation (Dataset2). We adopted per-concept average precision (AP) for evaluation based on manual groundtruth as well as mean AP (MAP) for all concepts.

5.1. Evaluation on Wearable Camera Images (Dataset1)

For this evaluation, we assess TFR method on the same dataset as in [26], indexed by a set of 85 everyday concepts with 12,248 images collected from 4 users with wearable cameras. To test the performance on different levels of concept detection accuracy, detectors were simulated using the *Monte Carlo* method following the work in [1]. In this simulation, concept detection performance is controlled by modifying the models' parameters based on manually annotated groundtruth of concept occurrences. These parameters are the mean μ_1 and standard deviation σ_1 for the positive class, as well as the mean μ_0 and the standard deviation σ_0 for the negative class. The performance of concept detection can be varied by controlling the intersection of the areas under the two probability density curves by changing the means or the standard deviations of the two classes for a single concept detector. During the simulation procedure, we fixed the two standard deviations and the mean of the negative class and varied the mean of the positive class μ_1 in the range [1.0...5.0], the original detection accuracy results for individual concepts are simulated and MAP is shown in Fig. 2 (denoted as Original) as semantic indexing results before refinement. Since the increasing of μ_1

reduced the intersection area of positive and negative class distributions, the original detection accuracy are improved accordingly as shown in Fig. 2.

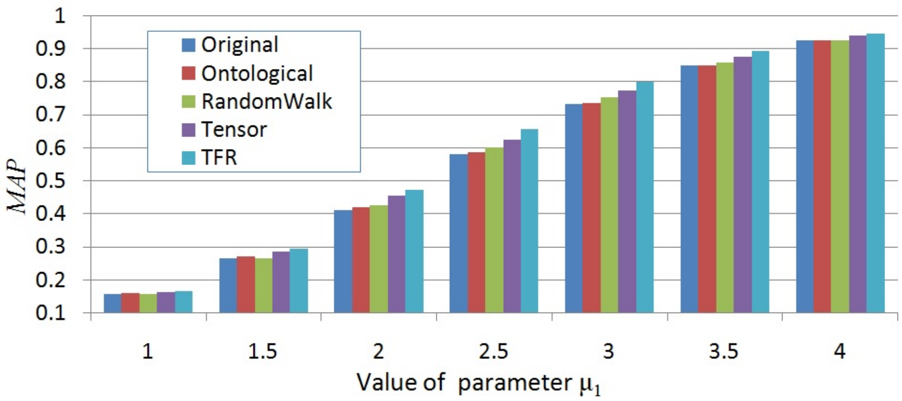

Figure 2: *MAP* of TFR refinement, Ontological, Random Walk, Tensor and Original on the wearable sensing dataset (mean over 20 runs)

In Fig. 2, the TFR method is compared with a variety of concept detection refinement methods including ontological refinement [28], a Random Walk-based method [24], as well as the state-of-the-art Tensor-based refinement for wearable sensing [26]. In ontological refinement, an ontology is constructed on 85 concepts with *subsumption* and *disjointness* concept relationships. 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 between classification accuracy and multi-class margin. The same ontology is also applied to TFR. Note that the ontology is not a pre-requisite to TFR as shown in Section 5.2 in which TFR can still achieve a comparable result to the state-of-the-art without an ontology and training step. To be fair, the Random Walk is performed in the same training-free manner, which means the concept co-occurrence is also inferred from thresholded pseudo-positive samples. The concept graph is then constructed with each weight representing concept co-occurrence similarities. The original confidence scores of concept detections are then adjusted by random walk algorithm which propagates the scores with concept graph. In Tensor-based refinement, a tensor is employed to formalize event segmentations and concept detections in order to preserve the temporal characteristics of each event. A weighted non-negative tensor factorization is then applied to re-estimate the concept detection confidences according to concept patterns [27]. In TFR, we empirically choose the number of latent features as $d = 10$ and we threshold the detection results with 0.3. The fusion parameter in Eqn. (2) is simply set to $\alpha = 0.5$, assigning equal importance to the two matrices. We also use 30 nearest neighbours in the propagation step.

As we can see, TFR out-performs all the other methods at all levels of original detection MAP from $0.15@\mu_1 = 1.0$ to $0.92@\mu_1 = 4.0$. At $\mu_1 = 1.0$, the less significant performance of all refinement approaches makes sense as initial detection accuracy is low. In this case, very few correctly detected concepts are selected for further enhancement which is impractical in real world applications and counter to our assumption of reliability (Sec. 3). When original detection performance is good, as shown in Fig. 2 if $\mu_1 \geq 4.0$, there is no space to improve detection accuracy. Therefore, the improvement is not that significant at $\mu_1 \geq 4.0$ for all refinements. However, TFR still achieves the best refinement in both extreme cases.

The best of the overall improvements of different approaches are shown in Table 1, in which the corresponding accuracy levels are depicted with

μ_1 values. As shown, TFR out-performs other approaches significantly and obtains the highest overall MAP improvement of 14.6%. Recall that Tensor-based refinement uses the temporal neighbourhood patterns within image sequences but is still out-performed by the TFR method. The number of improved concepts is shown in Table 1, counted from a per-concept AP comparison before and after refinement. TFR can improve the detection of almost all concepts (80 out of 85). Due to the constraints of the ontology model with its fixed lexicon, only a limited number of concepts can be refined in the ontological method (only 30 concepts are improved). However, this does not limit the TFR methods which exploit various semantics.

Table 1: Top overall performance of approaches to semantic refinement. Abbreviations of Onto, RW and Tens represent ontological refinement, Random Walk-based method and Tensor-based refinement respectively.

Method	Onto	RW	Tens	TFR
Top Impr	3.2%	3.9%	10.6%	**14.6%**
Num Impr	30	56	**80**	**80**
Accu level	$\mu_1 = 1.5$	$\mu_1 = 2.5$	$\mu_1 = 2.0$	$\mu_1 = 2.0$

5.2. Evaluation on TRECVid Video (Dataset2)

Experiments were also conducted in the domain of broadcast TV news to assess the generality of TFR using the TRECVid 2006 video dataset [6, 8]. Dataset2 contains 80 hours broadcast TV news video segmented into 79,484 shots in total. As a multi-concept detection task, in TRECVid 2006 the dataset is indexed by a lexicon of 374 LSCOM concepts [17] and 20 concepts are selected for performance evaluation with their groundtruth provided.

We employed the reported performance of the official evaluated concepts by VIREO-374 as a baseline[1], which is based on building SVM models of 374 LSCOM concepts [7]. The performance of TFR is also compared to the state-of-the-art domain adaptive semantic diffusion (DASD) [6] technique on the same 20 evaluated concepts by TRECVid using the official metric of $AP@2000$, as shown in Fig. 3.

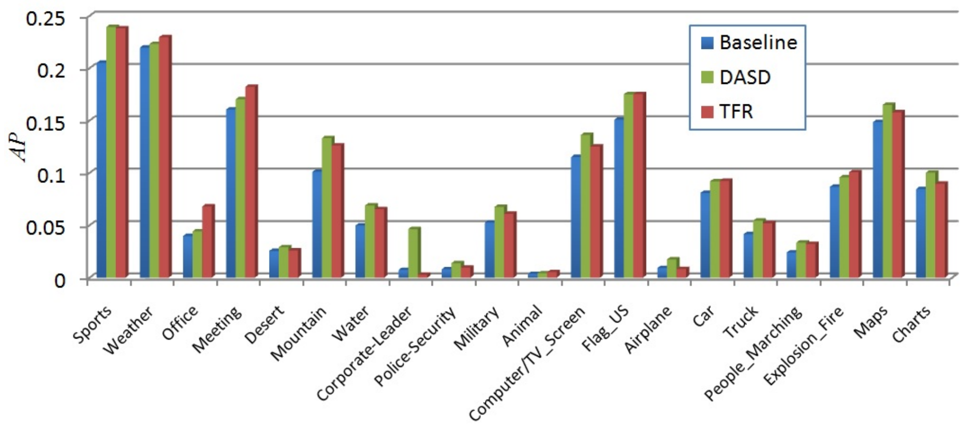

Figure 3: Per-concept $AP@2000$ comparison on the TRECVid 2006 dataset.

In our evaluation, TFR is implemented without using a concept ontology. The same parameters are applied directly as were used in Dataset1 without further optimization. As demonstrated, the results on Dataset2 are also promising using the same parameter values of d, α, etc., showing these parameters to be dataset independent. Similar as DASD, TFR achieves consistent enhancement gain against the baseline except for the concept of "Corporate_Leader", which is degraded in terms of performance. This is because "Corporate_Leader" only has 22 positive samples within the 79,484 samples in Dataset2, which makes accurately exploiting contextual patterns

[1] http://vireo.cs.cityu.edu.hk/research/vireo374/

from such few samples quite difficult. Over all other 19 concepts, the performance of TFR is comparable with DASD. Interestingly, according to our evaluation TFR does not require many positive samples in order to achieve satisfactory refinement. In Dataset2, the number of positive samples ranges from 150 to 1,556 and there are 10 of the 20 concepts which have less than 300 positive samples but still achieve satisfactory refinement by TFR. Note that DASD is still a training-based refinement method which needs to construct an initial concept semantic graph through learning from the TRECVid 2005 dataset whereas training data or *a priori* knowledge are not a pre-requisite for TFR.

5.3. The Effect of Different Semantics

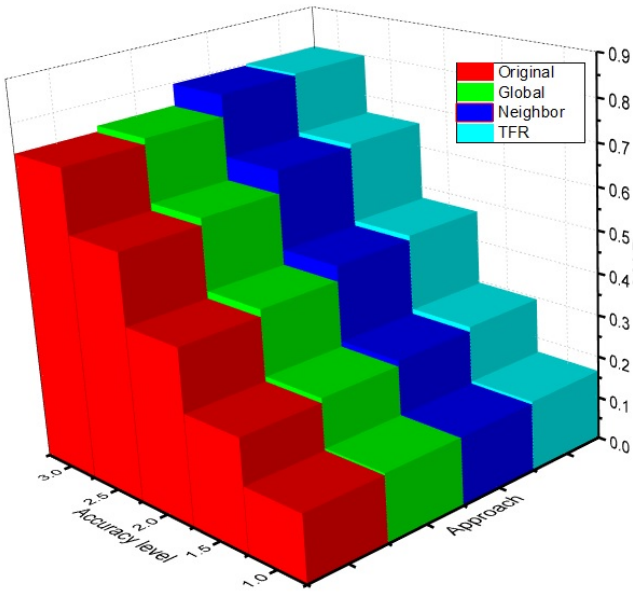

Figure 4: Effect comparison of different semantics in refinement. TFR obtains the highest by integrating them in a unified framework (Dataset1).

Fig. 4 depicts the roles of different semantics in refinement of semantic indexing at original detection accuracy levels of $\mu_1 = [1.0, ..., 3.0]$ in Dataset1. The *Global* in Fig. 4 is generated using an intermediate refined result C' with ontological weighting by GR. *Neighbour* is generated using the original C as input for neighbourhood-based propagation instead of using C'. While the exploitation of contextual and neighbourhood semantics can both refine the original indexing results, TFR can further integrate them to achieve the most significant refinement. Generally speaking, refinement by neighbourhood relationships will tend to adapt to the dataset better than global patterns, especially when original accuracy is high enough since the neighbours are more reliable and can better refine the target sample through similarity propagation in this case. Furthermore, by calculating the pair-wise similarity on the globally refined results C', the final results obtained by TFR are further improved. This is because the less accurate detections are first refined in C' hence will be less likely to disruptively affect the neighbourhood-based propagation.

As described in Section 4.1, reliable detection results can be selected by thresholding the original confidences for refining low-accuracy counterparts. The threshold indeed decides the number of trustworthy elements in C which can be used for context-based refinements. The number of reliable elements (depicted as density in C) and their correlation with the threshold is depicted in Table 2, for which the improvement is judged using the intermediate result C'. The density decreases while threshold value increases because fewer elements can be selected and regarded as accurate enough to carry out the refinement.

On the contrary, at a given detection accuracy level (fixed μ_1), the improve-

Table 2: Effect of reliable detections (Dataset1) evaluated on intermediate result C'.

	$\mu_1 = 1$		$\mu_1 = 2$		$\mu_1 = 3$	
thres	**Dens**	**Impr**	**Dens**	**Impr**	**Dens**	**Impr**
0.2	17.3%	**1.4%**	9.6 %	2.7%	7.7%	1.5%
0.3	10.4%	**1.4%**	7.3%	3.1%	6.8%	1.7%
0.4	6.5%	1.0%	5.8%	**3.2%**	6.1%	1.8%
0.5	4.1%	0.6%	4.7%	3.1%	5.7%	**1.9%**

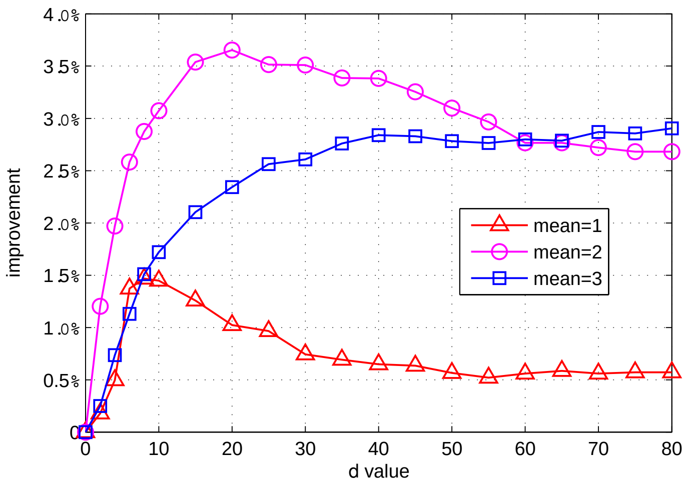

Figure 5: Impact of latent features (Dataset1) evaluated on intermediate result C'.

ment climbs first and then drops as the threshold increases continuously. This is because high/low thresholding criteria lead to insufficient/incorrect detections which are not reliable enough for refinement and this verifies the assumption of detection reliability as introduced in Section 3. The best performance is obtained when the threshold in the range $[0.3, 0.5]$ for different μ_1 values. As shown in Table 2, if the original concept detection performance improves (i.e., larger μ_1), a higher threshold can be assigned accordingly in order to achieve better overall semantic enhancement. This is because increasing the threshold will induce fewer misclassified concepts which are regarded as reliable, when the original detections are more accurate.

The impact of selected latent features is shown in Fig. 5 in which the MAP improvement is assessed on the intermediate result C' for $\mu_1 = 1$, 2 and 3, depicted across different d values. When original concept detection does not perform well, better improvement is achieved when fewer latent features are selected. This can be shown by the peaks at $d = 8$ and 20 for $\mu_1 = 1$ and 2 respectively. With the increase in d, the performance decreases gradually and converges at stable values. More stable performance is shown for better original detections such as at $\mu_1 = 3$ at which the performance keeps increasing and usually converges when about 40 latent features are selected. The small number d of latent features needed for refinement verifies the compactness assumption of projected semantic space which can be reconstructed with lower-rank dimensions, as introduced in Section 3.

The ontological weighting algorithm described in Section 4.2 was applied and incorporated with the WNMF-based enhancement to take advantage of the function of the ontology. In this experiment, we directly employed the same concept ontology structure as used in Section 5.1 and applied the concept semantics in choosing each weight element in matrix W to alleviate the deficiency introduced by uniform weighting. In Fig. 6, the ontological weighting approach is compared with the WNMF-based approach with uniform weighting scheme. As demonstrated in Fig. 6, the ontological weighting scheme significantly outperforms the uniform weighting scheme, which shows great potential for concept semantics if they are employed effectively in concept detection. The ontological weighting scheme combined with WNMF-based enhancement not only has better performance than the WNMF-based method, but also complements the shortcoming of WNMF-based enhancement at small μ_1 values. According to experiments, the WNMF-based method plus the ontological weighting scheme outperforms both of them over various concept detection accuracies.

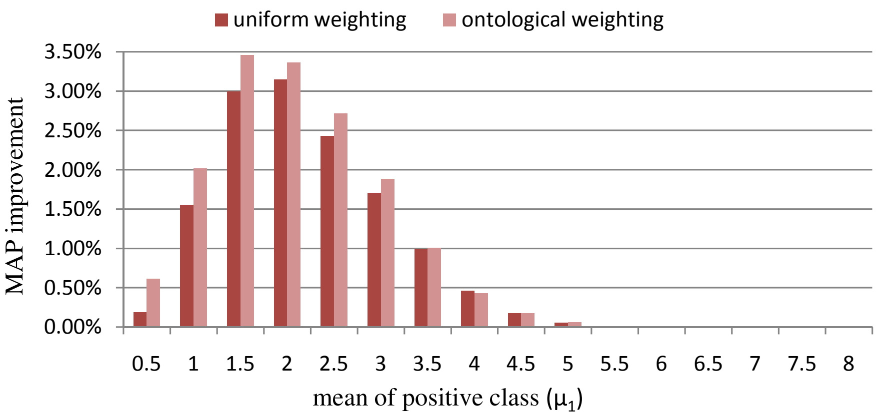

Figure 6: Improvement after using ontological weighting.

According to the above results, our TRF algorithm has many advantages. First, the approach is data-efficient and easy to implement. It can obtain significant detection enhancement even if there is no prior knowledge such as an ontology structure or distributions learned from extra training data. Second, the approach is shown to be effective in significantly improving detection accuracies for a large number of concepts. If combined with ontological weighting, the approach shows even better enhancement performance. Finally, the only input required are the initial concept detection results and the algorithm is independent of any specific implementation of concept detectors, the advantage of which is domain-independence.

5.4. Efficiency Analysis of TFR

In each iteration using Eqn. (8), the computational complexity is only relevant to the dimensionality of the matrix C and the selection of low rank d. For a total of $iter$ iterations to converge, the running time is thus $O(iter \cdot NMd^2)$. The complexity of TFR is linear to the size of concept lexicon. This can be easily scaled up to much larger concept lexicon and is more promising compared to learning models such as multi-label training whose complexity is quadratic to the number of concepts.

Recall that $d \leq min\{N, M\}$ and the number of concepts M in the lexicon is usually much smaller than the number of instances in the corpus N. Hence the computational complexity can be simplified as $O(iter \cdot N)$. In our experiments, the updating step of the approximation of L and R only takes several hundred iterations to obtain satisfactory approximation. Thus we empirically fix $iter = 1,000$ and for Dataset1, it takes approximately 30 seconds to execute the factorization on a conventional desktop computer.

Similarly, the computational complexity for graph propagation on one target sample can be represented as $O(iter \cdot kM * k^2)$. Since a small fixed value for k is enough in the implementation, the total complexity for neighbourhood-based refinement is also $O(iter \cdot N)$ which indicates the TFR method can be easily scaled up to much larger corpora.

6. Conclusions

Heterogenous multimedia content generated for various purposes usually have high visual and semantic diversities, thus presenting a barrier to the current approaches usually taken to refinement for concept-based semantic indexing, which highly depend on the quality of a training corpus. To ease these challenges, we presented the motivation for a training-free semantic refinement (TFR) of visual concepts, aimed at maximizing indexing accuracy by exploiting trustworthy annotations. TFR can take advantage of various semantics including global contextual patterns, ontologies or other knowledge structures and temporal neighbourhood relationships, all within a unified framework.

The rationale and algorithm presented in this paper have been assessed on two different datasets from very different domains and collected for very different applications, in order to show its versatility. Though exempt from the training/learning steps, the performance of TFR is still found to be comparable or better than the state-of-the-art. Since TFR is based on the assumption that reliable detection results can be selected as cues for refinement, a study of adaptive selection strategy is one area for future work. Besides traditional refinement tasks, TFR can also be applied in social tag recommendation, cross-domain label refinement, and others.

Acknowledgements

This work was part-funded by 973 Program under Grant No. 2011CB302206, National Natural Science Foundation of China under Grant No. 61272231, 61472204, 61502264, Beijing Key Laboratory of Networked Multimedia and by Science Foundation Ireland under grant SFI/12/RC/2289. We also thank Prof. Philip S. Yu for helpful discussions.

[1] R. Aly, D. Hiemstra, F. de Jong, P. Apers, Simulating the future of concept-based video retrieval under improved detector performance, Multimedia Tools and Applications (2011) 1–29.

[2] G. Awad, C.G.M. Snoek, A.F. Smeaton, G. Quénot, TRECVid Seman-

tic Indexing of Video: A 6-Year Retrospective, ITE Trans. on Media Technology and Applications (submitted) (2016).

[3] S. Ayache, G. Quénot, Video Corpus Annotation using Active Learning, in: European Conference on Information Retrieval (ECIR), Glasgow, Scotland, pp. 187–198.

[4] D. Brickley, R.V. Guha, RDF vocabulary description language 1.0: RDF Schema, W3C Technical Report (2004).

[5] W. Jiang, S.F. Chang, A. Loui, Context-based concept fusion with boosted conditional random fields, in: Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on, volume 1, pp. I–949–I–952.

[6] Y.G. Jiang, Q. Dai, J. Wang, C.W. Ngo, X. Xue, S.F. Chang, Fast semantic diffusion for large-scale context-based image and video annotation, Image Processing, IEEE Transactions on 21 (2012) 3080–3091.

[7] Y.G. Jiang, C.W. Ngo, J. Yang, Towards optimal bag-of-features for object categorization and semantic video retrieval, in: Proceedings of the 6th ACM International Conference on Image and Video Retrieval, CIVR '07, ACM, New York, NY, USA, 2007, pp. 494–501.

[8] Y.G. Jiang, J. Wang, S.F. Chang, C.W. Ngo, Domain adaptive semantic diffusion for large scale context-based video annotation, in: Computer Vision, 2009 IEEE 12th International Conference on, pp. 1420–1427.

[9] Y. Jin, L. Khan, L. Wang, M. Awad, Image annotations by combining multiple evidence & wordnet, in: Proceedings of the 13th Annual ACM International Conference on Multimedia, MULTIMEDIA '05, ACM, New York, NY, USA, 2005, pp. 706–715.

[10] L.S. Kennedy, S.F. Chang, A reranking approach for context-based concept fusion in video indexing and retrieval, in: Proceedings of the 6th ACM International Conference on Image and Video Retrieval, CIVR'07, ACM, New York, NY, USA, 2007, pp. 333–340.

[11] D.D. Lee, H.S. Seung, Algorithms for Non-negative Matrix Factorization, in: Advances in Neural Information Processing Systems, MIT Press, 2001, pp. 556–562.

[12] B. Li, K. Goh, E.Y. Chang, Confidence-based dynamic ensemble for image annotation and semantics discovery, in: Proceedings of the Eleventh ACM International Conference on Multimedia, MULTIMEDIA '03, ACM, New York, NY, USA, 2003, pp. 195–206.

[13] J. Liu, Y. Zhang, Z. Li, H. Lu, Correlation consistency constrained probabilistic matrix factorization for social tag refinement, Neurocomputing 119 (2013) 3 – 9. Intelligent Processing Techniques for Semantic-based Image and Video Retrieval.

[14] Z. Lu, L. Wang, J.R. Wen, Image classification by visual bag-of-words refinement and reduction, Neurocomputing 173, Part 2 (2016) 373 – 384.

[15] G.S. Mike Dean, OWL Web ontology language reference, W3C Recommendation (2004).

[16] D. Morrison, S. Marchand-Maillet, E. Bruno, Tagcaptcha: Annotating images with captchas, in: Proceedings of the ACM SIGKDD Workshop on Human Computation, HCOMP '09, ACM, New York, NY, USA, 2009, pp. 44–45.

[17] M. Naphade, J.R. Smith, J. Tesic, S.F. Chang, W. Hsu, L. Kennedy, A. Hauptmann, J. Curtis, Large-scale concept ontology for multimedia, IEEE Multimedia 13 (2006) 86–91.

[18] G.J. Qi, X.S. Hua, Y. Rui, J. Tang, T. Mei, H.J. Zhang, Correlative multi-label video annotation, in: Proceedings of the 15th International Conference on Multimedia, MULTIMEDIA '07, ACM, New York, NY, USA, 2007, pp. 17–26.

[19] B. Safadi, S. Ayache, G. Quénot, Active Cleaning for Video Corpus Annotation, in: MMM 2012 - International MultiMedia Modeling Conference, Klagenfurt, Austria, pp. 518–528.

[20] A. Smeaton, P. Over, 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, Springer, 2008, pp. 151–174.

[21] A.F. Smeaton, P. Over, W. Kraaij, Evaluation campaigns and TRECVid, in: Proceedings of the 8th ACM international workshop on Multimedia information retrieval, ACM, pp. 321–330.

[22] C.G.M. Snoek, M. Worring, Concept-based video retrieval, Foundations and Trends in Information Retrieval 2 (2008) 215–322.

[23] A. Ulges, C. Schulze, D. Keysers, T.M. Breuel, Computer vision systems: 6th international conference, icvs 2008 santorini, greece, may 12-15, 2008 proceedings, Computer Vision Systems: 6th International Conference, ICVS 2008 Santorini, Greece, May 12-15, 2008 Proceedings, Springer Berlin Heidelberg, Berlin, Heidelberg, 2008, pp. 415–424.

[24] C. Wang, F. Jing, L. Zhang, H.J. Zhang, Image annotation refinement using random walk with restarts, in: Proceedings of the 14th Annual ACM International Conference on Multimedia, MULTIMEDIA '06, ACM, New York, NY, USA, 2006, pp. 647–650.

[25] C. Wang, F. Jing, L. Zhang, H.J. Zhang, Content-based image annotation refinement, in: Computer Vision and Pattern Recognition, 2007. CVPR '07. IEEE Conference on, pp. 1–8.

[26] P. Wang, A. Smeaton, C. Gurrin, Factorizing time-aware multi-way tensors for enhancing semantic wearable sensing, in: 21st International Conference, MMM 2015, pp. 571–582.

[27] P. Wang, L. Sun, S. Yang, A.F. Smeaton, C. Gurrin, Characterizing everyday activities from visual lifelogs based on enhancing concept representation, Computer Vision and Image Understanding 148 (2016) 181 – 192.

[28] Y. Wu, B. Tseng, J. Smith, Ontology-based multi-classification learning for video concept detection, in: IEEE International Conference on Multimedia and Expo, volume 2, pp. 1003–1006 Vol.2.

[29] D. Xu, P. Cui, W. Zhu, S. Yang, Find you from your friends: Graph-based residence location prediction for users in social media, in: Multimedia and Expo (ICME), 2014 IEEE International Conference on, pp. 1–6.

[30] X. Xue, W. Zhang, J. Zhang, B. Wu, J. Fan, Y. Lu, Correlative multi-label multi-instance image annotation, in: Computer Vision (ICCV), 2011 IEEE International Conference on, pp. 651–658.

[31] Z.J. Zha, T. Mei, Y.T. Zheng, Z. Wang, X.S. Hua, A comprehensive representation scheme for video semantic ontology and its applications in semantic concept detection, Neurocomputing 95 (2012) 29 – 39. Learning from Social Media Network.