# Towards Training-Free Refinement for Semantic Indexing of Visual Media

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

National Laboratory for Information Science and Technology
Department of Computer Science and Technology, Tsinghua University
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.** 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 or post-processing the outputs of individual concept detection. Due to the limitations and availability of training corpora which are usually sparsely and imprecisely labeled, 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 global and temporal neighbourhood information inferred from the original concept detections in terms of weighted non-negative matrix factorization and neighbourhood-based graph propagation, respectively. Any available ontological concept relationships can also be integrated into this model as an additional source of external *a priori* knowledge. Experiments on two datasets 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 widespread in everyday scenarios. While mainstream consumer-based use of video such as YouTube and Vine are based on user

* 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.

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 abstract (e.g., *violence*, *meeting*). The conventional approach to content-based indexing, as taken in the annual TRECVid benchmarking [13, 12], is to annotate a collection covering both positive and negative examples, for the presence of each concept and then to train a machine learning classifier to recognise 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 and triples 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.

Context-Based Concept Fusion (CBCF) is an approach to refining the detection results for independent concepts by modeling relationships between them [2]. Concept correlations are either learned from annotation sets [7, 15, 16, 4, 3] or inferred from pre-constructed knowledge bases [18, 6] 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. 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. TFR can fully exploit global patterns of multi-concept appearance and an ontology (if available), as well as sampling the distribution of concept occurrences in the neighbourhood to enhance the original one-per-class concept detectors, all 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.

## 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, 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 [11], 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. Detection refinement has attracted interest based on exploiting concept correlations inferred from annotation sets [7, 15, 16, 2] or from pre-constructed knowledge bases [18, 6, 9]. 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 [6], sigmoid function [18], mutual information [7], random walk [15, 16], random field [2], 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 [4, 3], the concept graph is learned from the training set. The migration of concept alinement to testing sets, is based on the assumption of the homogeneity of two data sets, which is not always the case and can reduce the performance of indexing user-generated media, for example. The proposed TRF method in this paper is indeed a refinement methods but tries to tackle the above challenges.

## 3 Motivation and Proposed Solution

Fusing the results of concept detection to provid 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 [12, 14]. 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 relatated activities. Likewise still images of a social event captured in sequence will have semantic relationships based on shared locations, activities or people. For such "connected" visual media it makes sense to try to exploit any 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.

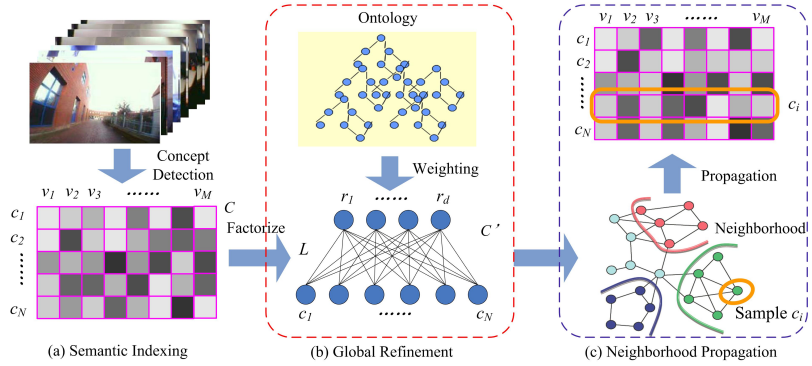


Fig. 1: Illustration of the TFR framework. (a) Semantic Indexing: Media samples indexed through concept detections, returning $C$. (b) Global Refinement (GR): Refining $C$ as $C'$ using global contextual patterns. (c) Neighbourhood propagation (NP): Refining $C'$ by similarity propagation between nearest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

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 $c_i (1 \leq i \leq N)$ represents a sample media element such as an

image or video shot, while each column corresponds to a concept $v_j (1 \leq j \leq M)$ in the vocabulary. We use different gray levels to represent matrix elements in $C$, namely the confidences of concept detections.

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. 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 $C$. Ontological relationships among concepts if they exist 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 $c_i$. 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.

The application of 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 $\tilde{c}_{ij} = \sum_{k=1}^{d} l_{ik} r_{kj}$. In GR, we optimize the factorization problem in weighted low rank to reflect different accuracies of concept detections. 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 [7, 17] under the same assumption that the initial detectors are reasonably reliable if the returned confidences are larger than a threshold. To distinguish contributions of different concept detectors to the cost function, we employ a weight matrix $W = (w_{ij})_{N\times M}$ whose elements are larger for reliable and lower for less reliable detections, and 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 [8] 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 [8], 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 [8]:

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

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

$$\nabla F(r^t) + K(r^t) r - K(r^t) r^t = 0 \tag{5}$$

where

$$\nabla F(r^t) = L^T D_w (L r^t - c) + \lambda r^t \tag{6}$$

The combination of Eqn. (5) and (6) achieves the update rule

$$R_{kj}^{t+1} \leftarrow R_{kj}^t \frac{[L^T (C \circ W)]_{kj}}{[L^T (LR \circ W)]_{kj} + \lambda R_{kj}} \tag{7}$$

where $\circ$ denotes Hadamard (element-wise) multiplication and each element in $L$ can be updated similarly. Note that it is not hard to prove convergence under the update rule of Eqn. (7) by proving $G(r, r^t)$ as an auxiliary function of $F$.

### 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. The goal is to correctly construct the matrix $W$ which can reflect the interaction between concepts and their detection accuracy. The confidence of sample $x$ belonging to concept $v$ being returned by a detector is represented as $Conf(v|x)$. We introduce the multi-class margin factor [9] as $Conf(v|x) - max_{v_i \in D} Conf(v_i|x)$, where $D$ is the universal set of disjoint concepts of $v$ which contains all concepts exclusively occurring with $v$.

By employing an ontology we assign each element in $W$ as

$$w_{ij} \propto 1 - [c_{ij} - max_{v_k \in D} c_{ik}] \tag{8}$$

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 $x_i$. 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 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.

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 $c_i$ and $c_j$, 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 $c'_i = (c'_{ik})_{1 \le k \le M}$ 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

$$P'_{i,j} = e^{-\frac{(1-P_{i,j})^2}{2\delta^2}} \tag{9}$$

where $\delta$ 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. (9).

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 [19] 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. (9). 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 $t_{i,j} = P'_{i,j} / \sum_{l=1}^{k+1} P'_{l,j}$, 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 $C_n^t \leftarrow TC_n^{t-1}$, 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 temporal 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 [17], 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]. By varying the controlling parameter $\mu_1$ in the range [1.0...5.0], the original detection accuracy results for individual concepts

is simulated and $MAP$ is shown in Fig. 2 (denoted as Original) as semantic indexing results before refinement.

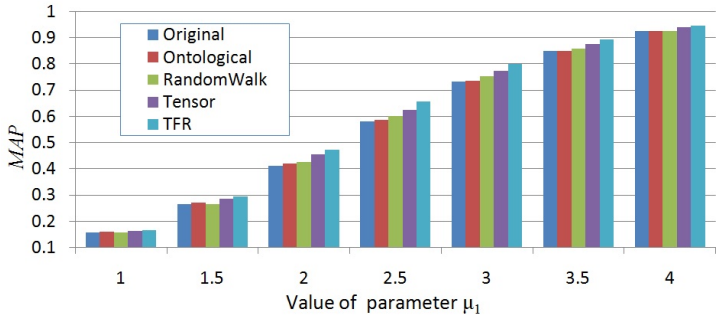


Fig. 2: $MAP$ of TFR refinement, Ontological, Random Walk, Tensor and Original on the wearable sensing dataset (mean over 20 runs)

In Fig. 2, the TFR method is compared with a variety of concept detection refinement methods including ontological refinement [18], a Random Walk-based method [15], as well as the state-of-the-art Tensor-based refinement for wearable sensing [17]. An ontology is constructed on 85 concepts with *subsumption* and *disjointness* concept relationships and 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. 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 $\mu_1$ 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.

| **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 [3, 4]. 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 [10] 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 [5]. The performance of TFR is also compared to the state-of-the-art domain adaptive semantic diffusion (DASD) [3] technique on the same 20 evaluated concepts by TRECVid using the official metric of $AP$@2000, as shown in Fig. 3.

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 $d$, $\alpha$, 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 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.

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

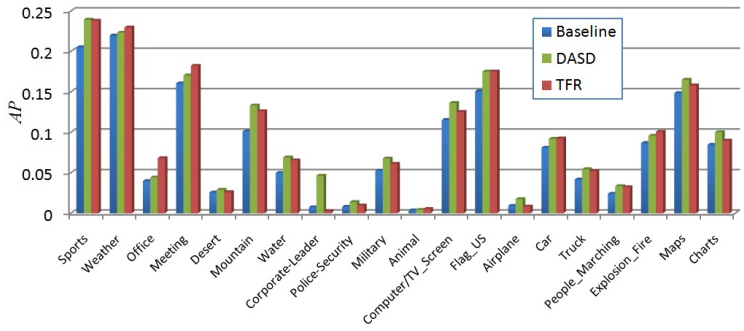


Fig. 3: Per-concept $AP$@2000 comparison on the TRECVid 2006 dataset.

### 5.3 Efficiency Analysis of TFR

In each iteration using Eqn. (7), 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 $O(iter \cdot NMd^2)$.

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. 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.

## References


1. R. Aly, D. Hiemstra, F. de Jong, and P. Apers. Simulating the future of concept-based video retrieval under improved detector performance. *Multimedia Tools and Applications*, 60(1):203–231, 2012.
2. W. Jiang, S.-F. Chang, and A. Loui. Context-based concept fusion with boosted conditional random fields. In *ICASSP*, pages I–949, 2007.
3. Y.-G. Jiang, Q. Dai, J. Wang, C.-W. Ngo, X. Xue, and S.-F. Chang. Fast semantic diffusion for large-scale context-based image and video annotation. *IEEE Trans. on Image Proc.*, 21(6):3080–3091, 2012.
4. Y.-G. Jiang, J. Wang, S.-F. Chang, and C.-W. Ngo. Domain adaptive semantic diffusion for large scale context-based video annotation. *ICCV*, 1420–1427, 2009.
5. Y.-G. Jiang, C.-W. Ngo, and J. Yang. Towards optimal bag-of-features for object categorization and semantic video retrieval. In *Proceedings of the 6th ACM Intnl. conference on Image and video retrieval (CIVR)*, 494–501. ACM, 2007.
6. Y. Jin, L. Khan, L. Wang, and M. Awad. Image annotations by combining multiple evidence & WordNet. In *ACM Multimedia*, pages 706–715, 2005.
7. L. S. Kennedy and S.-F. Chang. A reranking approach for context-based concept fusion in video indexing and retrieval. In *CIVR*, pages 333–340. ACM, 2007.
8. D. D. Lee and H. S. Seung. Algorithms for Non-negative Matrix Factorization. In *NIPS*, pages 556–562. MIT Press, Apr. 2001.
9. B. Li, K. Goh, and E. Y. Chang. Confidence-based dynamic ensemble for image annotation and semantics discovery. In *ACM Multimedia*, pages 195–206, 2003.
10. M. Naphade, J. R. Smith, J. Tesic, S.-F. Chang, W. Hsu, L. Kennedy, A. Hauptmann, and J. Curtis. Large-scale concept ontology for multimedia. *IEEE Multimedia*, 13(3):86–91, 2006.
11. G.-J. Qi, X.-S. Hua, Y. Rui, J. Tang, T. Mei, and H.-J. Zhang. Correlative multi-label video annotation. In *ACM Multimedia*, pages 17–26, 2007.
12. A. 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*, pages 151–174, 2008.
13. A. F. Smeaton, P. Over, and W. Kraaij. Evaluation campaigns and TRECVid. In *Proceedings of the ACM international workshop on Multimedia Information Retrieval*, pages 321–330. ACM, 2006.
14. C.G.M. Snoek and M. Worring. Concept-based video retrieval. *Foundations and Trends in Information Retrieval*, 2(4):215–322, 2008.
15. C. Wang, F. Jing, L. Zhang, and H.-J. Zhang. Image annotation refinement using random walk with restarts. In *ACM Multimedia*, pages 647–650, 2006.
16. C. Wang, F. Jing, L. Zhang, and H.-J. Zhang. Content-based image annotation refinement. In *CVPR*, pages 1–8, 2007.
17. P. Wang, A. F. Smeaton, and C. Gurrin. Factorizing time-aware multi-way tensors for enhancing semantic wearable sensing. In *MultiMedia Modeling, MMM*, volume 8935, pages 571–582. Springer LNCS, 2015.
18. Y. Wu, B. Tseng, and J. Smith. Ontology-based multi-classification learning for video concept detection. In *ICME*, pages 1003–1006, Vol.2, 2004.
19. D. Xu, P. Cui, W. Zhu, and S. Yang. Find you from your friends: Graph-based residence location prediction for users in social media. In *ICME*, pages 1–6, 2014.
20. X. Xue, W. Zhang, J. Zhang, B. Wu, J. Fan, and Y. Lu. Correlative multi-label multi-instance image annotation. In *ICCV*, pages 651–658. IEEE, 2011.