www.elsevier.com/locate/sigpro

High order pLSA for indexing tagged images

S. Nikolopoulos, S. Zafeiriou, I. Patras, I. Kompatsiaris

PII: S0165-1684(12)00271-X
DOI: http://dx.doi.org/10.1016/j.sigpro.2012.08.004
Reference: SIGPRO4783

To appear in: *Signal Processing*

Received date: 15 December 2011
Revised date: 16 July 2012
Accepted date: 3 August 2012

Cite this article as: S. Nikolopoulos, S. Zafeiriou, I. Patras and I. Kompatsiaris, High order pLSA for indexing tagged images, *Signal Processing,* http://dx.doi.org/10.1016/j.sigpro.2012.08.004

High Order pLSA for Indexing Tagged Images

S. Nikolopoulos[a,b,*], S. Zafeiriou[c], I. Patras[b], I. Kompatsiaris[a]

[a] *Centre for Research and Technology Hellas - Information Technologies Institute - 6th km Charilaou-Thermi Road, Thermi-Thessaloniki - GR-57001 Thessaloniki - Greece - Tel. +30-2311.257701-3, Fax.+30-2310-474128*

[b] *School of Electronic Engineering and Computer Science - Queen Mary University of London, E1 4NS, London, UK - Tel. +44 20 7882 7523, Fax. +44 20 7882 7997*

[c] *Department of Electrical and Electronic Engineering, Imperial College London, UK*

Abstract

This work presents a method for the efficient indexing of tagged images. Tagged images are a common resource of social networks and occupy a large portion of the social media stream. Their basic characteristic is the co-existence of two heterogeneous information modalities i.e. visual and tag, which refer to the same abstract meaning. This multi-modal nature of tagged images makes their efficient indexing a challenging task that apart from dealing with the heterogeneity of modalities, it needs to also exploit their complementary information capacity. Towards this objective, we propose the extension of probabilistic Latent Semantic Analysis to higher order, so as to become applicable for more than two observable variables. Then, by treating images, visual features and tags as the three observable variables of an aspect model, we learn a space of latent topics that incorporates the semantics of both visual and tag information. Our novelty is on using the cross-modal dependencies learned from a corpus of images to approximate the joint distribution of the observable variables. By penalizing the co-existence of visual content and tags that are known from experience to exhibit low dependency, we manage to filter out the effect of noisy content in the resulting latent space.

Keywords: high order pLSA, multi-modal analysis, tagged image indexing,

*Corresponding author. Tel: +30-2311257752, Fax:+30-2310474128

Email addresses: nikolopo@iti.gr (S. Nikolopoulos), s.zafeiriou@imperial.ac.uk (S. Zafeiriou), i.patras@eecs.qmul.ac.uk (I. Patras), ikom@iti.gr (I. Kompatsiaris)

cross modal dependencies, aspect models, social media

1. Introduction

Semantic image indexing has been recognized as a particularly valuable task for various applications of content consumption. Current literature has made considerable progress in this direction especially for uni-modal scenarios. However, it is generally accepted that multi-modal analysis has the potential to further improve this process, provided that the obstacles arising from the heterogeneous nature of different modalities can be overcome. This is based on the fact that independently of whether the different modalities act cumulatively or complementary, when combined, they encompass a higher amount of information that can be exploited to improve the efficiency of the performed task. Web 2.0 and social networks have primarily motivated this idea by making available plentiful user tagged images [1]. In addition, the significant improvement of automatic image annotation systems like [2, 3, 4, 5], further motivates multi-modal analysis since the automatically extracted labels can take the place of tags and facilitate semantic image indexing using both the visual and textual modalities.

The need to obtain a joint, unique representation of tagged images calls for techniques that will manage to handle the very different characteristics exhibited by the visual and tag information. This is true both in terms of the raw features' nature, i.e. sparse, high-dimensional tag co-occurrence vectors extracted from tag descriptions, compared to usually dense and low-dimensional descriptors extracted from visual content, as well as in terms of their semantic capacity, i.e. while abstract concepts like "freedom" are more easily described with text, ambiguous concepts like "rock" are more easily grounded using visual information. Based on the above, one can pursue a solution to the multi-modal indexing problem by defining a joint feature space where the projection of uni-modal features will yield a homogeneous and semantically enriched image representation.

The most trivial approach in this direction is to define a joint feature space

by concatenating the individual uni-modal features extracted from both modalities, also known as early fusion. However, by indiscriminately placing features extracted from different modalities into a common feature vector, the resulting space is likely to be dominated by one of the combined modalities or lose its semantic consistency. This was the reason that researchers turned into the statistical characteristics of the data to overcome these problems. For instance, [6] uses information theory and a maximum entropy model in order to integrate heterogeneous data into a unique feature space, [7] finds statistical independent modalities from raw features and applies super-kernel fusion to determine their optimal combination, [8] presents several cross-modal association approaches under the linear correlation model, while [9] rely on canonical correlation analysis to learn the cross-modal associations and use them for semantic image indexing.

The most recent approaches rely on the use of probabilistic Latent Semantic Analysis (pLSA) to facilitate the combination of heterogeneous modalities. The pLSA-based aspect or topic model is a method originally proposed in [10] that allows to map a high-dimensional word distribution vector to a lower-dimensional topic vector (also called aspect vector). This model assumes that the content depicted by every image can be expressed as a mixture of multiple topics and that the occurrences of words in this content is a result of the topic mixture. Thus, the latent layer of topics that is introduced between the image and the tag or visual words appearing in its content, acts as a feature space where both types of words can be combined meaningfully. Moreover, given that the goal of pLSA is to learn a set of latent topics that will act as bottleneck variables when predicting words, apart from handling the heterogeneity of multimodal sources, pLSA is also encouraged for discovering the hidden relations between images. Examples of pLSA-based approaches include [11] where pLSA is used to infer which visual patterns describe each concept, as well as [12] where Latent Dirichlet Allocation (LDA) [13] is used to model each image as the mixture of topics/object parts depicted in the image.

However, even if the space of latent topics can be considered to satisfy the

requirement of combining the words extracted from heterogeneous modalities without introducing any bias or rendering them meaningless, it still neglects the fact that, being different expressions of the same abstract meaning, there is a certain amount of dependance between the tag and visual words that appear together very frequently. This additional requirement motivates the employment of methods that will allow the cross-word dependencies to influence the nature of the extracted latent topics. In this context we examine the use of high order pLSA to improve the semantic capacity of the derived latent topics. High order pLSA is essentially the application of pLSA to more than two observable variables allowing the incorporation of different word types into the analysis process. We treat images, visual features and tags as the three observable variables of an aspect model and we manage to extract a set of latent topics that incorporate the semantics of both the visual and tag information space. The innovative aspect of our approach is the integration of cross-word dependencies into the update rules of high order pLSA, and specifically in the approximation of the joint distribution between images, visual features and tags. In this way, we succeed in devising a feature extraction scheme where the co-existence of two words that are known from experience to appear together rather frequently is more important in defining the latent topics, than the co-existence of two words that rarely appear together and are likely to be the result of noise. For estimating the cross-words dependencies between visual and tag words we introduce the concept of *word-profiles*. A *word-profile* is a vector representing the occurrence distribution of a word in a large corpus of images. Given that *word-profiles* are essentially binary vectors with dimensionality equal to the number of images in the corpus, their vector distance constitutes a natural way for measuring the dependency between the words of two different modalities.

Finally, our contribution lies also on proposing a distributed calculation model for high-order pLSA. This model benefits from the multi-core facilities offered by modern processors and renders the proposed approach applicable in large scale datasets, which is a crucial requirement when engineering an indexing scheme. The main advantages of this model, consists in reducing the total

computational cost by a factor that approximates the number of cores offered by the utilized processor, as well as in regulating the memory requirements of the algorithm independently of the dataset size.

The remaining of this work is structured as follows. In Section 2 we review the related literature, while in Section 3 we formulate image retrieval as a problem of defining a semantics sensitive feature space. Section 4 describes different approaches for defining a multi-modal feature space and presents our approach on how to apply high order pLSA using cross-word dependencies. A distributed calculation model for high order pLSA is presented in Section 5, showing ways to tackle the high computational and memory requirements of our method. Finally, our experimental findings are presented in Section 6 and concluding remarks are drawn in Section 7.

2. Related work

The multi-modal aspect that is intrinsic in social media prompted many researchers to investigate specialized methods for their multi-modal analysis. Among the existing works we identify the ones relying on the use of aspect or topic models [14] and the definition of a latent semantic space. For instance the authors of [15, 16] use a pLSA-based model to support multi-modal image retrieval in flickr, using both visual features and tags. They propose to extent the standard single-layer pLSA model to multiple layers by introducing not just a single layer of topics, but a hierarchy of topics. In this way they manage to effectively combine the heterogeneous information carried by the different modalities of an image. Similarly, pLSA is also the model adopted by the approach presented in [17] for multi-modal image retrieval. However in this case the authors propose an approach to capture the patterns between images (i.e. text words and visual words) using the EM algorithm to determine the hidden layers connecting them. Although the authors' goal is to exploit the interactions between the different modes when defining the latent space, they eventually implement a simplified model where they assume that a pair of different words

are conditionally independent given the respective image. This is different from our approach that uses the cross-modal dependencies learned from a corpus of images to approximate the joint distribution of the observable variables. A modified version of the pLSA framework is also adopted in [18] where the authors propose a visual language model for object categorization. In this model, probabilistic latent topic analysis is employed to capture the spatial dependencies of local image patches by considering that the neighboring visual words extracted from these patches are dependent on each other. Based on this assumption, the visual words of the neighboring patches are formulated as the observation variables of an aspect model and the EM algorithm is employed to solve the optimization problem. However, in order to obtain analytically tractable density estimation for their models, they consider unigram, bigram and trigram models of the neighboring patches, which essentially models how often two or three neighboring patches co-occur in the same image. Again, this is different from our approach where instead of relying to the co-occurrence of visual and tag words in each image, we use the cross-modal dependencies learned from a corpus of images to approximate the joint probability distribution. Finally, the use of aspect models is also the approach followed in [19] for performing tag ranking and image retrieval. The authors extend the model of Latent Dirichlet Allocation (LDA) [13] to a new topic model called regularized LDA, which models the interrelations between images and exploits both the statistics of tags and visual affinities. In this way, they enforce visually similar images to pick similar distributions over topics.

In a similar fashion the authors of [20] propose an approach for the multi-modal characterization of social media by combining text features (e.g. tags) with spatial knowledge (e.g. geotags). The proposed approach is based on multi-modal Bayesian models which allow to integrate spatial semantics of social media in well-formed, probabilistic manner. In [21] a dual cross-media relevance model (DCMRM) is proposed for automatic image annotation, which performs image annotation by maximizing the joint probability of images and words. The proposed dual model involves two types of relations, word-to-word

and word-to-image relations, both of which are estimated by using search techniques on the web data. Although using both the visual and textual modalities, the focus of this work is mainly on using web images and commercial search engines to estimate the joint probability of images and words, which is different from our approach where the joint probability distribution is estimated based on the co-occurrence of visual features and tags in a large corpus of tagged images. In [22] the authors present an effective method for multi-label image classification, where a multi-label classifier based on uni-modal features and an ensemble classifier based on bi-modal features are integrated into a joint classification model to perform semantic image annotation. In this work latent semantic indexing is used to obtain the correlations among different labels and incorporate them into the classification process, however in contrast to our work no mechanism is presented for estimating and using the cross-modal dependencies, i.e. dependencies between visual features and semantic labels. The concept of Flickr distance presented in [23] is another case that aims to exploit the visual and textual information that characterize social media. Flickr distance is a measure of the semantic relation between two concepts using their visual characteristics. The authors rely on the assumption that images about the same concept share similar appearance features and use images obtained from flickr to build a visual language model for each concept. Then, the Flickr distance between different concepts is measured by the square root of Jensen-Shannon (JS) divergence between the corresponding visual language models.

Improving the retrieval performance of tagged images has been also encountered as a problem of tag relevance learning, with the visual content serving as the driver of the learning process. In this direction the authors of [24, 25] rely on the intuition that if different persons label visually similar images using the same tags, these tags are likely to reflect the objective aspects of the visual content. Then, based on this intuition, they propose a neighbor voting algorithm for learning tag relevance by propagating common tags through the visual links introduced by visual similarity. Similarly, the work presented in [26] proposes the use of a multi-edge graph for discovering the tags associated

with the underlying semantic regions in the image. Each vertex in the graph is characterized with a unique image and the multiple edges between two vertices are defined by thresholding the pairwise similarities between the individual regions of the corresponding images. Then, based on the assumption that any two images with the same tag will be linked at least by the edge connecting the two regions corresponding to the concerned tag, the repetition of such pairwise connections in a fraction of the labeled images is used to infer a common "visual prototype". Tag relevance learning is also the problem addressed in [27], which aims at leaning an optimal combination of the multi-modality correlations and generate a ranking function for tag recommendation. In order to do this, the authors use each modality to generate a ranking feature, and then apply the Rankboost [28] algorithm to learn an optimal combination of these features.

Recently, there has been also an increasing interest on extending the aspect models to higher order through the use of Tensors [29]. Under this line of works we can mention the tag recommendation system presented in [30] that proposes a unified framework to model the three types of entities that exist in a social tagging system: users, items and tags. These data are represented by a 3-order tensor, on which latent semantic analysis and dimensionality reduction is performed using the Higher Order Singular Value Decomposition (HOSVD) technique [31]. The HOSVD decomposition is used also by the authors of [17] in order to decompose a 3-order tensor in which the first dimension is images, the second is visual words and the third is the text words. By applying the HOSVD decomposition on this 3-order tensor the authors aim to detect the underlying and latent structure of the images by mapping the original data into a lower dimensional space. Finally, a 3-order tensor is used also by the authors of [32] that propose an approach to capture the latent semantics of Web data. In order to do that the authors apply the PARAFAC decomposition [33] which can be considered as a multi-dimensional correspondent to the singular value decomposition of a matrix. In this case the extracted latent topics are used for the task of relevance ranking and producing fine-grained descriptions of Web data.

Finally, the manifold learning algorithms has also received a lot of research attention for the purpose of combining the various modalities carried in multimedia items. The authors of [34] present a novel approach for near-duplicate video retrieval where they propose a new algorithm for multiple feature hashing that makes use of the key-frame's manifold information when learning the necessary hash codes. In [35] the authors rely on the assumption that images reside on a low-dimensional sub-manifold and propose a geometrically motivated relevance feedback scheme, which is naturally conducted only on the image manifold in question rather than the total ambient space. Low-level visual features including color, texture, etc, are mapped into high-level semantic concepts using a Radial Basis Function (RBF) neural network that exploits the user interactions in query-by-example system. Identifying the manifold structure using a set of images has been also employed for face recognition in [36]. The authors of this work model the manifold structure using a nearest-neighbor graph which preserves the local structure of the image space. Then, each face image in the image space is mapped into a low-dimensional face subspace, which is characterized by a set of feature images, called Laplacianfaces. This sub-space exhibits high discrimination power and manages to decrease the error rates in face recognition. Finally, in the direction of exploiting the relations at the level of multimedia documents and at the level of media objects expressed by different modalities, the authors of [37] first construct a Laplacian media object space for media object representation of each modality and a multimedia documents semantic graph to learn the semantic correlations between documents. Then, the characteristics of media objects propagate along the semantic graph and a semantic space is constructed to perform cross-media retrieval.

3. Problem formulation

In order to index tagged images based on their semantic meaning we need to define a feature space where the distance between two images is proportional to their semantic affinity. To put this formally, given an image d, the set of

concepts depicted by this image $C_d = \{c_1, c_2, \ldots, c_{|C|}\}$, a representation $F_S^d = \{f_{s_1}, f_{s_2}, \ldots, f_{s_{|S|}}\}$ of the image in feature space S, the distance $dist(F_S^{d_i}, F_S^{d_j}) \geq 0$ between the representations of two images in S and a set of D images indexed based on their representations; we need to define the feature space S where $\forall d \in D$ the typical image retrieval process $Q(d_q, D) = rank_r(dist(F_S^{d_q}, F_S^{d_r}))$ returns a ranked list of all images in D such that when $dist(F_S^{d_q}, F_S^{d_i}) \leq dist(F_S^{d_q}, F_S^{d_j})$ it also stands that $|C_{d_q} \cap C_{d_i}| \geq |C_{d_q} \cap C_{d_j}|$. Thus, image retrieval is essentially a problem of defining a semantics sensitive feature space. In the following we describe different techniques for defining a feature space suitable for indexing tagged images.

4. Building a semantics sensitive space for tagged images

4.1. Codebook-based representation

One of the most popular approaches for image representation is based on defining a set of representative "words" (i.e. a Codebook $W = \{w_1, w_2, \ldots, w_{|W|}\}$), that are able to span a sufficiently large portion of the information space that they are used to describe. Then, based on this Codebook each image can be represented as an occurrence count histogram of the representative "words" in its content. The critical factor in this process is to define a highly expressive Codebook, so as to cover every potential instantiation of the image content. In the following we describe how the Codebook representation approach can be applied in the case of visual content and tags, as well as how to mix different Codebooks for obtaining a multi-modal image representation.

4.1.1. Visual codebook

In order to represent the visual information carried by an image using the aforementioned Codebook-based approach, we need to define the set of visual words that will act as the representative "words" of our information space. For the purposes of our work we have used the scheme adopted in [38] that consists of the following 3 steps: a) the Difference of Gaussian filter is applied on the

gray scale version of an image to detect a set of key-points and scales respectively, b) the Scale Invariant Feature Transformation (SIFT) [39] is computed over the local region defined by the key-point and scale, and c) a Visual Word Vocabulary (i.e. Codebook $V = \{v_1, v_2, \ldots, v_{|V|}\}$) [40] is created by applying the k-means algorithm to cluster in K clusters, the total amount of SIFT descriptors that have been extracted from all images. In cases where the memory and computational requirements of k-means are prohibitive for the full set of SIFT descriptors, the common practice is to sub-sample the collection of descriptors so as to create a clustering input with reasonable computational and memory requirements. Although this strategy is likely to result in a visual codebook of inferior quality, the impact on the quality of the resulting bag-of-words representations is in most cases marginal. Then, using the Codebook V we vector quantize the SIFT descriptor of each interest point against the set of representative visual words. This is done by mapping the SIFT descriptor to its closest visual word and increasing the corresponding word count. By doing this for all key-points found in an image, the resulting K-dimensional image representation is the occurrence count histogram of the visual "words" in its content, $F_V^d = \{f_{v_1}, f_{v_2}, \ldots, f_{v_{|V|}}\}$.

4.1.2. Tag codebook

A similar approach has been adopted for representing the tag information that accompanies an image using a tag Codebook. As in the previous case, we need to define the set of representative tag "words" that will manage to span a sufficiently large portion of the tag information space. However, in this case there is no need to employ clustering for determining which words should be included in the Tag Word Vocabulary (i.e. Codebook $T = \{t_1, t_2, \ldots, t_{|T|}\}$). Instead, from a large volume of utilized tags we need to select the ones with minimum level of noise and maximum usage by the users. For the purposes of our work we have used the Codebook construction process followed in [38]. More specifically, a large number of images are downloaded from flickr along with their accompanying tags. Among the total set of unique tags that have

been used by the users, there is a number of tags that appear more than 100 times. Many of these unique tags arise from spelling errors, while some of them are names etc, which are meaningless for general image annotation. Thus, all these tags that appear more than 100 times are checked against the WordNet Lexical Database [41] and after removing the non-existing ones, we end up with the final list of tags. Out of this final list we select the first N that have been used more frequently to form the tag Codebook. Eventually, we use this Codebook to obtain for each image a N-dimensional occurrence count histogram of the tag "words" in its content, $F_T^d = \{f_{t_1}, f_{t_2}, \ldots, f_{t_{|T|}}\}$.

4.1.3. Combining visual and tag codebooks

The most straightforward approach to produce a multi-modal image representation is to consider a combined Codebook composed by simply extending the list of representative visual-"words" with the list of representative tag-"words" (i.e. $VT = \{v_1, v_2, \ldots, v_{|V|}, t_1, t_2, \ldots, t_{|T|}\}$). In this case the generated image representation is essentially the concatenation of visual- and tag-based representations, which results in a $(K+N)$-dimensional occurrence count histogram, $F_{VT}^d = \{f_{vt_1}, f_{vt_2}, \ldots, f_{vt_{|V|+|T|}}\}$.

Moreover, apart from the simple concatenation we have also employed two additional approaches for combining the visual and tag-based features spaces. In the first case the distance between two images is calculated as the average of the distances between their visual and tag-based representation, which are calculated independently in each feature space and normalized to yield a value between [0,1]. In the second case, Canonical Correlation Analysis (CCA) [42] is employed to learn a basis of canonical components for both the visual and tag feature space, which define a sub-space that maximize the correlation between the two modalities. Using these components the original visual and tag-based representations are projected into the extracted sub-spaces and concatenated to form a joint image representation.

The major drawback of all aforementioned codebook-based combination approaches is that concatenation is performed between heterogeneous quantities.

This results in an non-uniform feature space that is unable to exploit the complementary effect of different modalities. Motivated by this fact, pLSA has been proposed to create a uniform space for the combination of different modalities.

4.2. Mixture of latent topics

pLSA aims at introducing a latent (i.e. unobservable) topic layer between two observable variables (i.e. images and words in our case). Let us denote $D = \{d_1, \ldots, d_{|D|}\}$ the set of images and $W = \{w_1, \ldots, w_{|W|}\}$ the set of words. The key idea is to map high-dimensional word occurrence count vectors, as the ones described in Section 4.1, to a lower dimensional representation in a so-called latent semantic space [10]. pLSA is based on a statistical model which has been called aspect model [14]. The aspect model is a latent variable model for co-occurrence data $n(d, w)$ (see Fig 1(a) for an example), which associates an unobserved class variable $z \in Z = \{z_1, \ldots, z_{|Z|}\}$ with each observation as shown in Fig. 1(b). Then, given that $P(w|d)$ is the conditional probability of words given images that can be obtained by performing row-wise normalization of $n(d, w)$, a joint probability model over the set of images D and the set of words W is defined by the mixture:

$$P(d, w) = P(d)P(w|d), \quad P(w|d) = \sum_{z \in Z} P(w|z)P(z|d) \tag{1}$$

where $P(d)$ denotes the probability of an image to be picked, $P(z|d)$ the probability of a topic given the current image, and $P(w|z)$ the probability of a word given a topic.

[Figure 1 about here.]

Once a topic mixture $P(z|d)$ is derived for an image d, we have a high-level representation of this image with less dimensions from the initial representation that was based on the co-occurrence of words. This is because we commonly choose the number of topics $|Z|$ to be much smaller than the number of words so as to act as bottleneck variables in predicting words. The resulting

$|Z|-$dimensional topic vectors can be used directly in an image retrieval setting, if we take the distance (e.g. L_1, Euclidean, cosine) between the topic vectors of two images to express their similarity.

4.2.1. Visual-based latent topics

In the visual information space, pLSA can be applied by considering the representative visual "words" of the visual codebook to constitute the second observable variable. Then, using the co-occurrence vectors between images and visual words $n(d, v)$, each image of D can be represented in the visual-based latent space using the following joint probability model:

$$P(d, v) = P(d)P(v|d), \quad P(v|d) = \sum_{zv \in ZV} P(w|zv)P(zv|d) \tag{2}$$

In the visual-based latent space $P(zv|d)$, the vector elements of each image representation denote the degree to which an image can be expressed using the corresponding visual based latent topics, $F^d_{ZV} = \{f_{zv_1}, f_{zv_2}, \ldots, f_{zv_{|ZV|}}\}$.

4.2.2. Tag-based latent topics

Similarly, in the tag information space pLSA can be applied by considering the representative tag "words" of the tag codebook to constitute the second observable variable. Then, using the tag-word co-occurrence vectors between images and tag words $n(d, t)$, each image of D can be represented in the tag-based latent space using the following joint probability model:

$$P(d, t) = P(d)P(t|d), \quad P(t|d) = \sum_{zt \in ZT} P(w|zt)P(zt|d) \tag{3}$$

In the tag-based latent space $P(zt|d)$, the vector elements of each image representation denote the degree to which an image can be expressed using the corresponding tag-based latent topics, $F^d_{ZT} = \{f_{zt_1}, f_{zt_2}, \ldots, f_{zt_{|ZT|}}\}$

4.2.3. Combining visual and tag based latent space

Motivated by the fact that both topic vectors refer to the so-called latent semantic space and express probabilities (i.e. the degree to which a certain topic

exists in the image), we assume that the topics obtained from both modalities are homogeneous and can be indiscriminately considered as the representative "words" of a combined codebook. Based on this assumption, an image representation that combines information from both modalities can be constructed by concatenating into a common multi-modal image representation, the two image representations of visual and tag based latent space, $F_Z^d = \{f_{zv_1}, f_{zv_2}, \ldots, f_{zv_{|ZV|}}, f_{zt_1}, f_{zt_2}, \ldots, f_{zt_{|ZT|}}\}$.

However, even if the concatenation is performed between values of similar nature (i.e. latent topics obtained through the application of pLSA), the simple combination of visual and tag based topics completely neglects the dependencies that may exist between the original visual- and tag-"words". Thus, even if we know by experience that the visual word v_i has low dependency with the tag word t_j, there is no way for the aforementioned approach to exploit this knowledge. This shortcoming was the basic motivation for applying high order pLSA as detailed subsequently.

4.3. High order pLSA

[Figure 2 about here.]

High order pLSA is the extension of pLSA to more than two observable variables. Using high order pLSA our goal is to apply the previously described aspect model for our three observable variables namely images, visual words and tag words. Using the asymmetric approach for pLSA, the generative model for our three observable variables is graphically represented in Fig 2 and can be expressed as follows:

$$P(d, v, t) = P(d) \sum_Z P(v|z)P(t|z)P(z|d) \tag{4}$$

In this case, if we introduce $R(z, v, t, d)$ to indicate which hidden topic z is selected to generate v and t in d such that $\sum_z R(z, v, t, d) = 1$, the complete likelihood can be formulated as:

$$L = \sum_D \sum_V \sum_T P(d,v,t) \sum_Z R(z,d,v,t) \\ [logP(d) + logP(v|z) + logP(t|z) + logP(z|d)] \quad (5)$$

and the function that we need to maximize is:

$$E[L] = \sum_D \sum_V \sum_T P(d,v,t) \sum_Z P(z|d,v,t) \\ [logP(d) + logP(v|z) + logP(t|z) + logP(z|d)] \quad (6)$$

Thus, using Expectation Maximization (EM) [43] the latent topics can be learned by randomly initializing $P(v|z)$, $P(t|z)$ and $P(z|d)$ and iterating through the following steps:

E-step:

$$P(z|d,v,t) = \frac{P(v|z)P(t|z)P(z|d)}{\sum_Z P(v|z)P(t|z)P(z|d)} \quad (7)$$

M-step:

$$P(d) = \frac{\sum_V \sum_T \sum_Z P(d,v,t)P(z|v,t,d)}{\sum_D \sum_V \sum_T \sum_Z P(d,v,t)P(z|v,t,d)}$$

$$P(v|z) = \frac{\sum_D \sum_T P(d,v,t)P(z|v,t,d)}{\sum_D \sum_T \sum_V P(d,v,t)P(z|v,t,d)}$$

$$P(t|z) = \frac{\sum_D \sum_V P(d,v,t)P(z|v,t,d)}{\sum_D \sum_T \sum_V P(d,v,t)P(z|v,t,d)}$$

$$P(z|d) = \frac{\sum_V \sum_T P(d,v,t)P(z|v,t,d)}{\sum_Z \sum_V \sum_T P(d,v,t)P(z|v,t,d)} \quad (8)$$

whereas for indexing a new image I_q we just need to repeat the above steps but without updating $P(d)$, $P(v|z)$ and $P(t|z)$ that have been obtained from the learning stage. The iterations stop when the value of eq.(6) converge to its maximum (either local or global). The convergence properties of the algorithm have been proven in [44] and [45]. As convergence criterion we use the relative change of $E[L]$ between consecutive iterations, as shown in eq.(9). If this relative change is below a predefined threshold $thre$ the process is terminated, otherwise the EM steps are repeated.

$$\frac{E_{current}[L] - E_{previous}[L]}{abs(E_{previous}[L])} = \begin{cases} \geq thre, & \text{repeat} \\ < thre, & \text{terminate} \end{cases} \tag{9}$$

In eqs.(5-8) we have used the joint probability distribution $P(d, v, t)$ of the observable variables (i.e. documents, visual words and tag words), in order to formulate high order pLSA. Due to the normalizing denominators, instead of $P(d, v, t)$ any un-normalized approximation to it can be used. The classical pLSA formulations use the frequency of occurrence $n(d, v, t)$, which is the number of times a visual word v_i appears together with a tag word t_j in a given image d_k. However, in our effort to incorporate prior knowledge into the generation process of the latent topics, we have followed an approach where $P(d, v, t)$ is approximated using the cross-word dependencies. More specifically, we accept that there is a certain degree of dependence on how visual words appear together with tag words, and that this dependence can be learned from data. In order to estimate these dependencies we introduce the concept of *word-profiles*. The *word-profile* is a $|D|$-dimensional binary vector that models the occurrence distribution of a *word* in a set of $|D|$ images, having 1's in the places corresponding to the images where the *word* appears at least once and 0 in all other places. In other words, the *word-profiles* are the column vectors of $n(d, t)$ and $n(d, v)$ after thresholding them with 1. Using the occurrence distribution of each *word* in a corpus of images, we have a natural way to estimate the dependency between *words* of different type (i.e. visual and tag), by measuring their vector distance. For the purposes of our work, given that the values of *word-*

profiles cannot be negative, we have used the complement of cosine similarity to calculate the dependency between two words v and t. In a set of preliminary experiments the jaccard distance and mutual information were also tested for quantifying the dependency between v and t. However, the cosine similarity metric was found to deliver the latent space with the best retrieval performance and was favored for our experiments. Thus, the cross-word dependency between v and t is calculated as shown below.

$$J(v,t) = 1 - \frac{v * t}{\parallel v \parallel \parallel t \parallel} \tag{10}$$

Then, $\forall v \in V$ and $\forall t \in T$ we calculate $J(v,t)$ in order to measure the dependency degree of every possible combination between the visual and tag words. Finally, we incorporate this information during the approximation of $P(d,v,t)$ as follows:

$$P(d,v,t) = \overline{n}(d,v) * \overline{n}(d,t) * J(v,t) \tag{11}$$

where $\overline{n}(d,v)$ and $\overline{n}(d,t)$ are the matrices $n(d,v)$, $n(d,t)$ after thresholding. The rationale behind using eq.(11) to approximate $P(d,v,t)$ is to penalize or favor the contribution of some pair (v,t) to the sum of eq.(8), based on the prior knowledge that we have about the dependency of v with t. In this way, the co-existence of a pair (v,t) with high cross-word dependency is more important in defining the mixture of latent topics, than the co-existence of a pair with low cross-word dependency, which can be the result of noise. In the remaining of the manuscript, including the result tables presented in Section 6, all references to high-order pLSA imply the version of the model described in eqs.(4-8) using the eqs.(10) and (11) to approximate $P(d,v,t)$.

5. A distributed model for calculating high-order pLSA

Although flexible for incorporating two or more random variables in a single latent space, high-order pLSA comes at the price of particularly high computational and memory requirements. As illustrated in eqs.(7-8) the algorithmic

implementation of high order pLSA will have to store in memory and traverse one 4-dimensional array for executing the update steps of EM. Given that the dimensionality of the codebook-based representation in both tag and visual space can range from a few hundreds to a few thousands, it is obvious that the resulting 4-dimensional matrix will become difficult to handle when the number of considered images becomes high. Although data sparseness can be used to alleviate this burden, still the high dimensionality of the matrices that need to be processed renders the proposed approach intractable for very large datasets.

Motivated by this fact, we propose a distributed calculation model for high-order pLSA that could benefit from the multi-core facilities offered by modern processors. Drawing from the literature in distributed clustering [46] and in analogy with the approach presented in [47] for distributed pLSA, we divide the full set of images into equally sized nodes. Each of these nodes is able to apply the algorithm locally and periodically communicate with a central super-node in order to synchronize with the other nodes. More specifically, using the notation of Section 4.3, the algorithm proceeds as follows:

1. Initially, the normalized, term-document co-occurrence matrices $P(d, v)$ and $P(d, t)$ are split along the images dimension into equally sized chunks $P^i(d, v)$ and $P^i(d, t)$. Every chunk is then transmitted to one of the K nodes so that each node carries the information for $|D|/K$ images, except the last node that may have less.
2. The super-node initializes with random values the matrices $P(v|z)$, $P(t|z)$, $P(z|d)$ and with equal priors the matrix $P(d)$. A copy of the matrices $P(v|z)$ and $P(t|z)$ is transmitted in all K nodes, while the matrices $P(d)$ and $P(z|d)$ are split along the images dimension into equally sized chunks $P^i(d)$ and $P^i(z|d)$, in order to be transmitted to each of the K nodes.
3. Each node calculates the local joint probability distribution $P^i(z|d, v, t)$ according to eq.(7) and estimates the local value $E^i[L]$ according to eq.(6). Then, the super-node sums the $E^i[L]$ values collected from all nodes in order to calculate the central $E[L]$ value for this iteration.

4. Each node locally calculates $P^i(d, v, t)$ based on eq.(11), by using $P^i(d, v)$ and $P^i(d, t)$, as well as the cross-words dependencies $J(v, t)$ that are common for all nodes.
5. After calculating $P^i(d, v, t)$ each node locally proceeds to the maximization step and produces the local matrices $P^i(d)$, $P^i(v|z)$, $P^i(t|z)$ and $P^i(z|d)$. The only difference from eq.(8) is that all 4 matrices are un-normalized (i.e. all denominators in eq.(8) are set to 1).
6. The local matrices $P^i(v|z)$ and $P^i(t|z)$ are collected from all nodes. The super-node performs element wise summation across i and normalizes the resulting matrices so that each column sum to 1. In this way the super-node updates the values of the global matrices $P(v|z)$ and $P(t|z)$, which are once again transmitted to all nodes.
7. Using the updated global matrices $P(v|z)$ and $P(t|z)$ and the corresponding $P^i(d)$ and $P^i(z|d)$ each node re-calculates the new local joint probability distribution $\acute{P}^i(z|d, v, t)$ according to eq.(7) and estimates the new local value of $\acute{E}^i[L]$ according to eq.(6). As in step 4, the super node sums the $\acute{E}^i[L]$ values collected from all nodes in order to calculate the new central $\acute{E}[L]$ value.
8. Using $\acute{E}[L]$ and $E[L]$ the super-node checks whether the convergence criterion of eq. (9) is satisfied. If yes, the local matrices $P^i(z|d)$ from all nodes are collected and concatenated in order to re-assemble the global matrix $P(z|d)$. If not, the process continues with Step 4.

By adopting this model for the distributed calculation of high order pLSA the benefit is twofold. Firstly, the fact that there is no need for communication or concurrent memory access between the nodes, allows them to run in parallel and synchronize only when they need to communicate with the super node. This parallel computation allow us to expect a reduction of the total computational time by a factor that approximates the number of cores offered by the utilized processor. Secondly, the proposed distributed model provides an elegant way for regulating the memory requirements of the algorithm independently of the

dataset size. Indeed, given that in a non-parallel mode the minimum amount of data that should be loaded into RAM is bounded by $P^i(z|d,v,t)$ instead of $P(z|d,v,t)$, allow us to implement a version of the model that fits the memory specifications of the utilized computer. This can be done by using more nodes with smaller size or vice versa. In section 6.3.4 of our experimental study we measure the gain in computational cost of the distributed calculation model and show how we can regulate our algorithm to process a significantly large set of images.

Finally, we should mention that apart from dealing with computational and memory limitations, the distributed calculation model is also suggested for applications where data sources are distributed over a network and collecting all data at a central location is not a viable option. These applications include privacy-preserving environments where each node is only allowed to share a sub-set or an encoded representation of the local data, as well as sensor networks where each node collects a set of observations and needs to design local processing rules that perform at least as well as global ones, which rely on all observations being centrally available.

6. Experimental Evaluation

Our experimental evaluation is primarily focused on comparing the performance achieved by the different feature spaces described in Section 4, in an image retrieval setting. Our aim is to verify that by exploiting the multi-modal nature of tagged images and introducing the cross-word dependencies when performing the modality fusion, we succeed in defining a feature space that is more sensitive to semantics. We also verify the efficiency of our approach in handling tasks of varying requirements by evaluating its performance in an image clustering setting. Moreover, we experimentally measure the gain in computational cost achieved by the distributed calculation model and show how we can significantly reduce the memory requirements of our algorithm and run high order pLSA on a significantly large set of images. Finally, we compare our work

with two state-of-the-art approaches that are also oriented towards exploiting the multi-modal nature of tagged images for improving the performance of an image retrieval system.

6.1. Data set

To carry out our evaluation we have used the NUS_WIDE[1] and the SOCIAL20 dataset[2]. The NUS_WIDE dataset was created by the NUS's Lab for Media Search [38] and contains 269,648 images that have been downloaded from flickr together with their tags. The NUS_WIDE dataset was selected due to its appropriateness for evaluating the examined retrieval and clustering tasks, its origin in social media (i.e. flickr) and its ability to facilitate large scale experiments. For all images we have used the 500-dimensional co-occurrence vectors for visual words and the 1000-dimensional co-occurrence vectors for tag-words that were released by the authors. Although sections 4.1.1 and 4.1.2 provide a few details about the uni-modal feature extraction process, additional information about this process as well as the statistical characteristics of the dataset (i.e. frequency distribution of tags, tags per image, etc) can be found in [38]. The ground-truth for all images with respect to 81 concepts has been provided to facilitate evaluation. The full set of 269,648 images has been split by the authors to 161,789 train and 107,859 test images. In our initial set of experiments that involved tuning the algorithm and observing its behavior against specific parameter we have used a sub-sample of 5,000 (I^{train}) images for training and 5,000 (I^{test}) images for testing (cf. Sections 6.3.1-6.3.4). Then, using the distributed calculation model described in Section 5 we have applied our approach on the full scale of the NUS_WIDE dataset, which constitutes a realistic configuration for an indexing framework (cf. Section 6.3.5).

On the other hand, the SOCIAL20 dataset consists of 19,971 images downloaded from flickr along with their tags. The ground-truth information contains

[1] http://lms.comp.nus.edu.sg/research/NUS-WIDE.htm

[2] http://staff.science.uva.nl/~xirong/index.php?n=DataSet.Social20

annotations with respect to 20 visually diverse concepts for the full set of images. The SOCIAL20 dataset was selected due to its origin in social media (i.e. flickr) and the nature of its ground-truth information consisting of visually diverse concepts. The employed uni-modal feature extraction process has been identical with the one used in the NUS_WIDE dataset, however in this case we have constructed more extended vocabularies resulting in 2000-dimensional feature vectors for both the visual and tag information space. The reason for doing so was on the one hand to demonstrate that the proposed distributed calculation model can efficiently deal with high dimensional image representations, and on the other hand to verify that high-order pLSA can deliver improvements with respect to baselines that employ large size vocabularies. Out of the $19,971$ images, 13314 have been used for training and 6657 for testing, corresponding to the 2/3 and 1/3 of the total dataset, respectively.

In both datasets, in order to remove the effects of incomplete tagging and noisy annotation, we have applied the restriction that all utilized images should have at least one concept present in their annotation info and at least one tag present in their tag-based representation.

6.2. Evaluation protocol

The adopted evaluation protocol is implemented as follows. Initially a set of training images is utilized to learn all necessary parameters that require training, such as the latent topics of simple and high order pLSA, as well as to calculate the cross-words dependencies between the visual and tag words. Subsequently, using the learned parameters the training images are indexed. Finally, an independent set of testing images is used to query the index and evaluate the system's performance based on the relevance between the query and the retrieved images.

For assessing the relevance between the query and the retrieved images we have used the Average Precision (AP) metric. AP favors the algorithms that are able not only to retrieve the correct images, but to retrieve them as early as possible in a ranked list of results. This is a crucial attribute for an image

retrieval system since users rarely take the time to browse through the results beyond the first pages. Average precision is expressed by the following equation.

$$AP = \frac{\sum_{r=1}^{N} Pr(r) \cdot rel(r)}{\#\text{ relevant images}} \tag{12}$$

where r is the current rank, N is the number of retrieved images, $rel()$ is a binary function that determines the relevance of the image at the given rank with the query image. $rel()$ outputs 1 if the image in the given rank is annotated with at least one concept in common with the query image and 0 otherwise. $Pr(r)$ is the precision at rank r and is calculated by:

$$Pr(r) = \frac{\#\text{ relevant retrieved images of rank r or less}}{r} \tag{13}$$

AP measures the retrieval performance of the method using one image as query. Finally, in order to facilitate fast image matching the images were indexed using a kd-tree multidimensional indexing structure [48] that supports k-NN (Nearest Neighbor) queries.

Apart from the AP and in order to evaluate the efficiency of our algorithm in a task different from image retrieval, we have also implemented the Normalized Mutual Information (NMI) measure for clustering comparison. NMI belongs to the class of information theoretic based measures that rely on the mutual information shared between two random variables. The mutual information measures how much knowing one of these variables reduces the uncertainty about the other, which makes it appropriate for measuring the similarity between two clustering solutions. The NMI measure that we have used in our work is a normalized version of the Mutual Information defined as:

$$NMI(U, V) = \frac{I(U, V)}{\sqrt{H(U)H(V)}} \tag{14}$$

where $I(\cdot)$ calculates the Mutual Information between the clustering solutions U and V and $H(\cdot)$ calculates the information entropy of each solution. NMI takes the value of 1 when the two clustering solutions are identical and 0 when they are independent. More information about NMI can be found in [49].

6.3. Results

6.3.1. Retrieval performance

In order to obtain one global performance score for all query images we employed the Mean Average Precision (MAP) score, which is the mean of AP scores over the full set of query images. In our experiments we have set the value of N to be equal with the total number of indexed images. As baseline we have used the performance scores obtained using the 8 different feature spaces described in Section 4 namely *visual-words*, *tag-words*, *visualtag-words*, *avgnorm-visualtags-words*, *cca-visualtag-words*, *plsavisual-words*, *plsatag-words* and *plsavisual_plsatag-words*. The performance score for the proposed approach appears under *highOrder-plsa*. The number of topics in all cases involving aspect models was selected to be 30, except from the *plsavisual_plsatag-words* case where the concatenation of the uni-modal plsa models resulted in a dimensionality of 60 topics. Moreover, in order to counterbalance the effect of initial randomization all experiments involving aspect models were repeated 5 times to obtain an average performance value. Table 1 shows the MAP scores for all evaluated feature spaces and for both the NUS_WIDE and SOCIAL20 datasets.

With respect to the NUS_WIDE dataset, we notice that *visual-words* performs better than *tag-words*. As expected, the straightforward combination of both modalities by simply concatenating their word count vectors *visualtag-words*, fails to combine them efficiently and performs slightly better than the visual modality, which seems to be the modality that dominates the joint feature space. However, the situation does not get any better when using the *avgnorm-visualtag-words* or the *cca-visualtag-words* approach to combine the uni-modal feature spaces. The fact that the performance scores in both cases reside very close to the scores achieved by the uni-modal feature spaces, shows that the heterogeneity of the feature spaces renders their efficient combination a challenging task. When moving to the space of pLSA-based latent topics we can see an increase of the retrieval performance for both uni-modal cases, which verifies the efficiency of aspect models to discover semantic relations between

the images. Moreover, it is interesting to note that the relative improvement achieved by *plsatag-words* is considerably higher than the relative improvement of *plsavisual-words*. This can be attributed to the ability of pLSA in more efficiently handling sparse data, since the co-occurrence table of tag-words is much more sparse than the corresponding table of visual words. Additionally, the performance achieved by *plsavisual_plsatag-words* introduces some improvement over the uni-modal cases, in contrast to the behavior of *visualtag-words*. This verifies the ability of the latent space to more efficiently combine the heterogeneous modalities of tagged images, compared to the original space of word counts. Finally, the performance achieved by the proposed method verifies the usefulness of cross-word dependencies in creating a semantics sensitive feature space. Indeed, we can see that *highOrder-plsa* outperforms all other cases that neglect this kind of dependencies, introducing an improvement of approximately 1.8% units over the best performing baseline.

With respect to the SOCIAL20 dataset, we notice a big difference between the performance achieved by the visual and textual modality, i.e. *tagwords* outperform *visualwords* by a significant amount. As in the previous case, the joint feature space resulting from the straightforward combination of both modalities is dominated from the visual modality, since the performance score of *visualtag-words* is identical with the performance of *visual-words*. The *cca-visualtag-words* seems to be heavily affected by the heterogeneity of the visual and tag-based representations since the obtained performance score is lower than the worse of the uni-modal cases. On the contrary, the situation gets better when using the *avgnorm-visualtag-words*, where by normalizing and averaging the independently calculated distances we succeed in devising a feature space that is not dominated by one of the modalities. However, the improvement is only marginal compared to the best performing uni-modal case since we fail to exploit the complementary information capacity of both modalities. In a fashion similar to the previous dataset the pLSA-based uni-modal feature spaces (i.e. *plstag-words* and *plsavisual-words*) deliver improvements with respect to their corresponding baselines, while the additional improvement introduced by their straightforward

concatenation shows again that the space of latent topics is more appropriate for combining the two different modalities. Finally, we notice that *highOrder-pLSA* significantly improves the retrieval performance of the resulting latent space, introducing an increase of approximately $24,8$ units over the best performing baseline. The fact that the images in SOCIAL20 are annotated with strictly one concept per image and that the 20 concepts utilized for annotation are selected based on their visually diversity[3], seems to particularly favor the employment of cross-modal dependencies in constructing a feature space sensitive to the semantics of these concepts.

[Table 1 about here.]

In order to gain more insight into the retrieval performance of our system we have calculated the MAP on a concept basis. In order to do this, for each concept, we have used only the images depicting this concept to query the index. Then, the MAP score of this concept is calculated by averaging the AP scores obtained for each of the issued queries. Figs. 3 and 4 depict the MAP scores achieved by the *plsavisual_plsatag-words* and *highOrder-plsa* approaches for the 30 concepts that appear more frequently in the NUS_WIDE test set and the 9 concepts that appear with a statistically safe frequency (more than 100 times) in the SOCIAL20 dataset. We can see that the proposed *highOrder-plsa* approach outperforms the best performing baseline in 21 out of the 30 concepts considered from the NUS_WIDE dataset and in all 9 concepts considered from the SOCIAL20 dataset.

[Figure 3 about here.]

[Figure 4 about here.]

Figs. 5, 6 are two illustrative examples that demonstrate the effect of aspect models and the potential of multi-modal analysis in the performance of image

[3]http://staff.science.uva.nl/∼xirong/index.php?n=DataSet.Social20

retrieval. Fig. 5 shows in ranked order the first 10 images retrieved using *tag-words*, *plsatag-words* and *highOrder-plsa*. With the query image depicting a *rhino*, the retrieval system manages to correctly retrieve various pictures of rhinos in the first 10 results. However, when using *tag-words* we can see that the system retrieves 2 outliers in the first 10 results, with the first of them being placed at rank #2 and the second at rank #8. The situation improves when using *plsatag-words* where the outliers are now placed at rank #7 and #9. Finally, the *highOrder-plsa* approach verifies the complementary information capacity of visual and tag words since there is only one outlier among the first 10 retrieved images which is placed at rank #6. In a similar fashion Fig. 6 shows in ranked order the first 10 images retrieved using *visual-words*, *plsavisual-words* and *highOrder-plsa*. With the query image depicting the concept *kitchen* all three feature spaces manage to retrieve pictures of a kitchen scene. However, as verified by the outliers existing at rank #4, #5, #7 and #10, *visual-words* relying solely on the image visual features tend to confuse the kitchen with the classroom scene. Using *plsavisual-words* the system reduces the number of outliers to 3 which are now placed at rank #3, #4 and #6, however it is not until we employ the *highOrder-plsa* space before we succeed in completely removing the outliers from the ranked list of 10 results. It is evident in these examples that the multi-modal nature of tagged images provides a solid ground for discovering their semantic relations and that the proposed approach is efficient in doing so.

[Figure 5 about here.]

[Figure 6 about here.]

6.3.2. Clustering Performance

In order to examine whether the proposed approach for semantic image indexing is appropriate for tasks other than image retrieval, we have designed an experiment where the task was to mine the conceptual categories characterizing a set of images. Relying on the fact that the authors of the NUS_WIDE dataset [38] provide a concept list where each of the 81 annotation concepts is classi-

fied to one of six categories namely *Event/Activities*, *Program*, *Scene/Location*, *People*, *Objects* and *Graphics*, the examined task was to automatically identify these categories by performing clustering on the images included in our test set I^{test}. In each case one of the aforementioned feature spaces was used for calculating the distance similarity matrix. Then, NMI was employed to compare each of the obtained clustering solutions against the solution derived from the ground truth information. The L1-norm metric was used to calculate the similarity distance between images and the k-means algorithm was employed to perform clustering. In all cases, the number of requested clusters was set to be equal with the number of categories and 100 repetitions were imposed on the clustering process in order to alleviate the sensitivity of k-means to the initial conditions. The obtained results are depicted in Table 2.

[Table 2 about here.]

It is evident from the NMI scores that the tag information space is more efficient in identifying the existing categories. Indeed, the clustering solutions obtained using *tag-words* and *plsatag-words* are much closer to the optimal solution than using *visual-words* and *plsavisual-words*, respectively. The poor performance of the visual information space is also observed in the cases of *visualtag-words* and *plsavisual_plsatag-words*, where the inclusion of visual-words in a joint space with tags has a negative effect on the clustering efficiency of the resulting space. This is also the case for *avgnorm-visualtag-words* and *cca-visualtag-words* where the performance scores observed, are very close to the *visual-words* baseline. Nevertheless, the use of cross-word dependencies by *highOrder-plsa* allows the resulting space to filter out the misleading information of visual words and obtain a clustering solution that is closer to the optimal case than all other baselines. This is an additional argument for the efficiency of the proposed approach in handling tasks of varying requirements.

6.3.3. Latent space dimensionality and convergence threshold

In this section we present the experiments performed on the sub-sample of the NUS_WIDE dataset in order to observed the behavior of our algorithm against two parameters, the dimensionality of the latent space and the convergence threshold employed for terminating the EM steps.

In the first case we investigate the impact of the dimensionality employed for the latent space on the retrieval performance of the pLSA-based methods. Our interest is on roughly estimating the number of dimensions where a performance peak is exhibited by each of the examined cases. Fig. 7 plots the MAP scores achieved by each method against the dimensionality of the latent space. We can see that the performance peak for *highOrder-plsa* appears between the range of $15-30$ dimensions. A similar kind of behavior is also exhibited by the uni-modal aspect models (i.e. *plsatag-words* and *plsavisual-words*) where the performance peak is located around the 30 dimensions. However, this is not the case for *plsavisual_plsatag-words* where the number of latent topics will have to reach 60 before achieving the peak of its performance. Thus, the proposed approach reaches its performance peak using considerably fewer dimensions than the best performing baseline. This fact constitutes an additional advantage of our approach since the efficiency of the indexing mechanisms, which are typically employed in image retrieval systems, benefit substantially from the low dimensionality of the utilized feature space.

[Figure 7 about here.]

In the second case, we examine the relation between the convergence threshold employed during the EM procedure and the retrieval performance of the resulting feature space. As already mentioned in Section 4.3 the iterations of the EM algorithm stop when the value of eq.(6) becomes lower than a predefined threshold. In all experiments so far this threshold was set to 10^{-3}. Here, we evaluate the retrieval performance of the proposed approach using as convergence threshold the values 10^{-4}, 10^{-5} and 10^{-6}. Fig. 8 shows the MAP scores for each of the aforementioned values. It is evident that by making the

convergence criterion more strict the retrieval performance of the resulting latent space increases. However, for values that are very close to zero (e.g. 10^{-5} and 10^{-6}) the improvement is only marginal and does not compensates for the increased computational overhead.

[Figure 8 about here.]

6.3.4. Distributed calculation model

In order to estimate the gain in computational cost achieved by the proposed distributed calculation model, we have used the sub-sample of the NUS_WIDE dataset and measured the time required by our high-Order pLSA algorithm to complete on an i7-950 processor with 4 physical cores and 12GBs of RAM, using the centralized and the distributed calculation model respectively. Moreover, for the distributed case we have considered two different configurations. In the first configuration titled "Distributed Memory", we consider that the memory facilities of the utilized computer are adequate to load in RAM the 4-dimensional $P(z|d, v, t)$ array that derives from the processed dataset, while in the second case titled "Distributed (Disk)", we consider that the memory required to load the $P(z|d, v, t)$ array exceeds the available resources. In this case the hard disk is used by each node to store and load the corresponding chunk $P^i(z|d, v, t)$ in every iteration. Table 3 demonstrates our experimental findings. We can see that the time required by our algorithm to complete reduces by a factor of ≈ 4 when employing the memory-based distributed calculation model, which is a reasonable outcome given that the whole process has been parallelized in 4 physical cores. On the other hand, when employing the configuration of the algorithm using the hard disk, the computational overhead introduced by read/write operations doubles the execution time but still remains considerably lower than the centralized version. Table 3 also depicts the computational cost required by some of the baseline algorithms presented in Section 6.3.1 that involve the calculation of an aspect model, namely *plsatag-words*, *plsavisual-words* and *plsavisual_plsatag-words*. It is evident that since these algorithms do not require the calculation of high-order models, the execution time is significantly

lower. However, the gain in performance compensates for the computational overhead.

[Table 3 about here.]

6.3.5. High-Order plsa in large scale

By exploiting the ability of the distributed calculation model to regulate its memory requirements, we have managed to apply the proposed high order pLSA algorithm to the full set of images provided by the NUS_WIDE dataset. More specifically, we have applied high order pLSA on $121,920$ train and $81,589$ test images, which is the set that constitutes the full NUS_WIDE dataset after removing the images that did not satisfy the restrictions described in Section 6.1. Table 4 shows the obtained MAP scores for all features space examined in Section 6.3.1, apart from the *avg-norm-visualtag-words* case that can not be supported by our kd-tree indexing infrastructure making the corresponding experiment extremely time-consuming. It is interesting to note that the improvements observed when moving from one feature space to another are equivalent to those observed in Table 1, advocating the effectiveness of our method to increase the semantic capacity of the resulting space, independently of the dataset scale.

[Table 4 about here.]

6.4. Comparison with existing methods

In order to compare our work with state-of-the-art methods in multi-modal indexing we have generated three additional feature spaces by implementing the methods proposed in [16], [17] and [18]. More specifically, we have implemented one of the variations presented in [16] that treats the visual and tag-based latent topics obtained from the application of the uni-modal pLSA, as the observed words for learning a second level pLSA model. This model (*ml-plsa* [16]) allows the image to be represented as a vector of meta-topics as illustrated in Fig. 9. Similarly, we have also implemented the multi-modal pLSA scheme (*mm-plsa*) presented in [17]. In this work the authors' goal is to exploit the interactions

between the different modes when defining the latent space. However, in order to simplify their model, they assume that the pair of random variables representing the visual and tag words are conditional independent, given the respective image d_i. Given this assumption, we have $P(v|t,d) = P(v|d)$ and the joint probability model of text words, visual words and images can be written as:

$$P(d,v,t) = P(d)P(t|d)P(v|t,d) \Rightarrow P(d,v,t) = P(d)P(t|d)P(v|d) \quad (15)$$

Finally, we have also adopted the approach presented in [18] for approximating the joint probability distribution. In this work, the unigram, bigram and trigram models adopted by the authors, suggest that the probability density estimations can be approximated by the co-occurrence counts of the observable variables. If we adopt this approach in our case, we have $P(v|t,d) = P(v \bigodot t|d)$, with $\bigodot$ counting how many times v_j appears together with t_i in d_k. Then, in this case, the joint probability model of text words, visual words and images can be written as:

$$P(d,v,t) = P(d)P(t|d)P(v|t,d) \Rightarrow P(d,v,t) = P(d)P(t|d)P(v \bigodot t|d) \quad (16)$$

Given eqs. 15 and 16 we have used the EM-steps of Section 4.3 to generate a feature space for the *mm-plsa* and the *count-plsa* model, respectively. Both of these models are different from our approach presented in Section 4.3, since in our case $P(d,v,t)$ is approximated using the cross-word dependencies (c.f. eq. 11). Table 5 compares the performance of the three methods obtained using both the sub-sample of the NUS_WIDE dataset, as well as the SOCIAL20 dataset.

[Figure 9 about here.]

[Table 5 about here.]

The fact that in both datasets the performance of *mm-plsa* model is lower than two of the baselines presented in Section 6.3.1 shows that there is important information neglected under the cross words independence assumption, and

the approximation of $P(d, v, t)$ without using the cross-words dependencies is misleading in the generation of a semantics sensitive latent space. This is further advocated by the fact that *count-plsa*, which also neglects the cross-word dependencies, is once again outperformed by the best performing baselines of Section 6.3.1. On the other hand, the *ml-plsa* model manages to introduce some improvement over the best performing baseline of Section 6.3.1. However, the improvement is marginal showing the the second level pLSA has little to offer when applied on dense data (i.e. such as the data produced by the application of the first level of pLSA). Finally, the fact that *highOrder-plsa* outperforms all other methods in both datasets shows that using the cross-word dependencies to approximate the joint distribution of the observable variables, is a promising direction towards combining the semantics of both visual and tag information space in a semantics sensitive latent space.

7. Conclusions

In conclusion, we should stress the great potential of exploiting the information carried by the different modalities of tagged images when designing a semantics sensitive feature space. The use of aspect models has proven to be an efficient solution for overcoming the heterogeneity of sources, allowing the resulting feature space to benefit from their complementary information capacity. Moreover, our experiments have shown that, being different representations of the same abstract meaning, the visual and tag words appearing in the image content exhibit some cross-word dependencies that can be used to improve the effectiveness of the resulting feature space. We have shown how the use of high order pLSA can be used to incorporate this type of dependencies and lead to performance improvements. In addition, by implementing a distributed model for the calculation of high order pLSA, we have shown how to benefit from multi-core facilities offered by modern processors and how to regulate the memory requirements of our algorithm so as to become applicable for datasets of very large size.

Moreover, the fact that the resulting latent space is semantically enhanced with information from both visual and tag information space renders the proposed approach appropriate for supporting various different tasks in social media consumption. For instance, it can be used to support tag-based image search, where tag relevance learning is achieved through visual content similarity [24, 25], or even diverse image search, where the retrieval mechanism is devised to ensure that the ranked list of results will consist of both relevant and diverse images [50] in the top places. Finally, it is important to note that although the approach presented in this work performs fusion between the modalities of visual features and tags, a similar methodology can be used to incorporate additional modalities of social media such as geo-located or user-related information. Increasing the number of considered modalities will be the main focus of our future work.

[1] C. Marlow, M. Naaman, D. Boyd, M. Davis, Ht06, tagging paper, taxonomy, flickr, academic article, to read, in: Hypertext, 2006, pp. 31–40.

[2] Trec video retrieval evaluation notebook papers and slides (Nov. 2011). URL http://www-nlpir.nist.gov/projects/tvpubs/tv.pubs.11.org.html

[3] M. Wang, X.-S. Hua, R. Hong, J. Tang, G.-J. Qi, Y. Song, Unified video annotation via multigraph learning, IEEE Transactions on Circuits and Systems for Video Technology 19 (5) (2009) 733 –746.

[4] M. Wang, X.-S. Hua, J. Tang, R. Hong, Beyond distance measurement: Constructing neighborhood similarity for video annotation, IEEE Transactions on Multimedia 11 (3) (2009) 465 –476.

[5] M. Wang, X.-S. Hua, T. Mei, J. Tang, G.-J. Qi, Y. Song, L.-R. Dai, Interactive video annotation by multi-concept multi-modality active learning, Int. J. Semantic Computing 1 (4) (2007) 459–477.

[6] J. Magalhaes, S. Rüger, Information-theoretic semantic multimedia indexing, in: CIVR '07, ACM, New York, USA, 2007, pp. 619–626.

[7] Y. Wu, E. Y. Chang, K. C.-C. Chang, J. R. Smith, Optimal multimodal fusion for multimedia data analysis, in: MULTIMEDIA '04, ACM, New York, USA, 2004, pp. 572–579.

[8] D. Li, N. Dimitrova, M. Li, I. K. Sethi, Multimedia content processing through cross-modal association, in: MULTIMEDIA '03, ACM, New York, USA, 2003, pp. 604–611.

[9] N. Rasiwasia, J. Costa Pereira, E. Coviello, G. Doyle, G. R. Lanckriet, R. Levy, N. Vasconcelos, A new approach to cross-modal multimedia retrieval, in: Proceedings of the international conference on Multimedia, MM '10, ACM, New York, NY, USA, 2010, pp. 251–260.

[10] T. Hofmann, Probabilistic latent semantic analysis, in: Proc. of Uncertainty in Artificial Intelligence, UAI'99, Stockholm, 1999.

[11] R. Lienhart, M. Slaney, Plsa on large scale image databases, in: Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on, Vol. 4, 2007, pp. IV–1217 –IV–1220.

[12] E. Hörster, R. Lienhart, M. Slaney, Image retrieval on large-scale image databases, in: Proceedings of the 6th ACM international conference on Image and video retrieval, CIVR '07, 2007, pp. 17–24.

[13] D. M. Blei, A. Y. Ng, M. I. Jordan, Latent dirichlet allocation, J. Mach. Learn. Res. 3 (2003) 993–1022.

[14] T. Hofmann, Unsupervised learning from dyadic data, MIT Press, 1998, pp. 466–472.

[15] S. Romberg, E. Hörster, R. Lienhart, Multimodal plsa on visual features and tags, in: Proceedings of the 2009 IEEE international conference on Multimedia and Expo, ICME'09, IEEE Press, 2009, pp. 414–417.

[16] R. Lienhart, S. Romberg, E. Hörster, Multilayer plsa for multimodal image retrieval, in: CIVR '09: Proceeding of the ACM International Conference on Image and Video Retrieval, ACM, New York, NY, USA, 2009, pp. 1–8.

[17] C. Pulla, C. V. Jawahar, Multi modal semantic indexing for image retrieval, in: Conference on Image and Video Retrieval, 2010, pp. 342–349.

[18] L. Wu, N. Yu, J. Liu, M. Li, Latent topic visual language model for object categorization, in: SIGMAP, 2011, pp. 149–158.

[19] H. Xu, J. Wang, X.-S. Hua, S. Li, Tag refinement by regularized lda, in: Proceedings of the 17th ACM international conference on Multimedia, MM '09, ACM, New York, NY, USA, 2009, pp. 573–576.

[20] S. Sizov, Geofolk: latent spatial semantics in web 2.0 social media, in: WSDM '10: Proceedings of the third ACM international conference on Web search and data mining, ACM, 2010, pp. 281–290.

[21] J. Liu, B. Wang, M. Li, Z. Li, W. Ma, H. Lu, S. Ma, Dual cross-media relevance model for image annotation, in: Proceedings of the 15th international conference on Multimedia, MULTIMEDIA '07, ACM, New York, NY, USA, 2007, pp. 605–614.

[22] W. Li, M. Sun, C. Habel, Multi-modal multi-label semantic indexing of images based on hybrid ensemble learning, in: Proceedings of the multimedia 8th Pacific Rim conference on Advances in multimedia information processing, PCM'07, Springer-Verlag, Berlin, Heidelberg, 2007, pp. 744–754.

[23] L. Wu, X.-S. Hua, N. Yu, W.-Y. Ma, S. Li, Flickr distance, in: ACM Multimedia, 2008, pp. 31–40.

[24] X. Li, C. G. Snoek, M. Worring, Learning tag relevance by neighbor voting for social image retrieval, in: Proceeding of the 1st ACM international conference on Multimedia information retrieval, MIR '08, ACM, New York, NY, USA, 2008, pp. 180–187.

[25] X. Li, C. G. M. Snoek, M. Worring, Learning social tag relevance by neighbor voting, IEEE Transactions on Multimedia 11 (2009) 1310–1322.

[26] D. Liu, S. Yan, Y. Rui, H.-J. Zhang, Unified tag analysis with multi-edge graph, in: Proceedings of the international conference on Multimedia, MM '10, ACM, New York, NY, USA, 2010, pp. 25–34.

[27] L. Wu, L. Yang, N. Yu, X.-S. Hua, Learning to tag, in: Proceedings of the 18th international conference on World wide web, WWW '09, ACM, New York, NY, USA, 2009, pp. 361–370.

[28] Y. Freund, R. Iyer, R. E. Schapire, Y. Singer, An efficient boosting algorithm for combining preferences, J. Mach. Learn. Res. 4 (2003) 933–969.

[29] T. G. Kolda, B. W. Bader, Tensor decompositions and applications, SIAM Review 51 (3) (2009) 455–500.

[30] P. Symeonidis, A. Nanopoulos, Y. Manolopoulos, Tag recommendations based on tensor dimensionality reduction, in: RecSys '08: Proceedings of the 2008 ACM conference on Recommender systems, ACM, New York, NY, USA, 2008, pp. 43–50.

[31] L. D. Lathauwer, B. D. Moor, J. Vandewalle, A multilinear singular value decomposition, SIAM J. Matrix Anal. Appl. 21 (4) (2000) 1253–1278.

[32] T. Franz, A. Schultz, S. Sizov, S. Staab, Triplerank: Ranking semantic web data by tensor decomposition, in: ISWC '09: Proceedings of the 8th International Semantic Web Conference, Springer-Verlag, Berlin, Heidelberg, 2009, pp. 213–228.

[33] R. A. Harshman, M. E. Lundy, Parafac: Parallel factor analysis, Computational Statistics & Data Analysis 18 (1) (1994) 39 – 72.

[34] J. Song, Y. Yang, Z. Huang, H. T. Shen, R. Hong, Multiple feature hashing for real-time large scale near-duplicate video retrieval, in: Proceedings of the 19th ACM international conference on Multimedia, MM '11, ACM, New York, NY, USA, 2011, pp. 423–432.

[35] X. He, W.-Y. Ma, H.-J. Zhang, Learning an image manifold for retrieval, in: Proceedings of the 12th annual ACM international conference on Multimedia, MULTIMEDIA '04, ACM, New York, NY, USA, 2004, pp. 17–23.

[36] X. He, S. Yan, Y. Hu, P. Niyogi, H.-J. Zhang, Face recognition using laplacianfaces, IEEE Transactions on Pattern Analysis and Machine Intelligence 27 (3) (2005) 328 –340.

[37] Y. Yang, Y.-T. Zhuang, F. Wu, Y.-H. Pan, Harmonizing hierarchical manifolds for multimedia document semantics understanding and cross-media retrieval, IEEE Transactions on Multimedia 10 (3) (2008) 437 –446.

[38] T.-S. Chua, J. Tang, R. Hong, H. Li, Z. Luo, Y. Zheng, Nus-wide: a real-world web image database from national university of singapore, in: CIVR '09: Proceeding of the ACM International Conference on Image and Video Retrieval, ACM, New York, NY, USA, 2009, pp. 1–9.

[39] D. G. Lowe, Distinctive image features from scale-invariant keypoints, Int. J. Comput. Vision 60 (2) (2004) 91–110.

[40] J. Sivic, A. Zisserman, Video google: A text retrieval approach to object matching in videos, in: ICCV '03: Proceedings of the Ninth IEEE International Conference on Computer Vision, IEEE Computer Society, Washington, DC, USA, 2003, p. 1470.

[41] C. Fellbaum (Ed.), WordNet: An Electronic Lexical Database (Language, Speech, and Communication), The MIT Press, 1998.

[42] R. A. Johnson, D. W. Wichern, Applied Multivariate Statistical Analysis (6th Edition), Prentice Hall, 2007.

[43] G. J. McLachlan, T. Krishnan, The EM algorithm and extensions, 2nd Edition, John Wiley and Sons, 1997.

[44] S. Zafeiriou, M. Petrou, Nonnegative tensor factorization as an alternative csiszar—tusnady procedure: algorithms, convergence, probabilistic inter-

pretations and novel probabilistic tensor latent variable analysis algorithms, Data Min. Knowl. Discov. 22 (3) (2011) 419–466.

[45] E. C. Chi, T. G. Kolda, On tensors, sparsity, and nonnegative factorizations, arXiv:1112.2414 [math.NA] (December 2011).
URL http://arxiv.org/abs/1112.2414

[46] S. Bandyopadhyay, C. Giannella, U. Maulik, H. Kargupta, K. Liu, S. Datta, Clustering distributed data streams in peer-to-peer environments, Information Sciences 176 (14) (2006) 1952 – 1985.

[47] B. Mehta, Learning from what others know: Privacy preserving cross system personalization, in: Proceedings of the 11th international conference on User Modeling, UM '07, Springer-Verlag, Berlin, Heidelberg, 2007, pp. 57–66.

[48] J. L. Bentley, Multidimensional binary search trees used for associative searching, Commun. ACM 18 (9) (1975) 509–517.

[49] N. X. Vinh, J. Epps, J. Bailey, Information theoretic measures for clusterings comparison: is a correction for chance necessary?, in: Proceedings of the 26th Annual International Conference on Machine Learning, ICML '09, ACM, New York, NY, USA, 2009, pp. 1073–1080.

[50] M. Wang, K. Yang, X.-S. Hua, H.-J. Zhang, Towards a relevant and diverse search of social images, IEEE Transactions on Multimedia 12 (8) (2010) 829 –842.

List of Figures

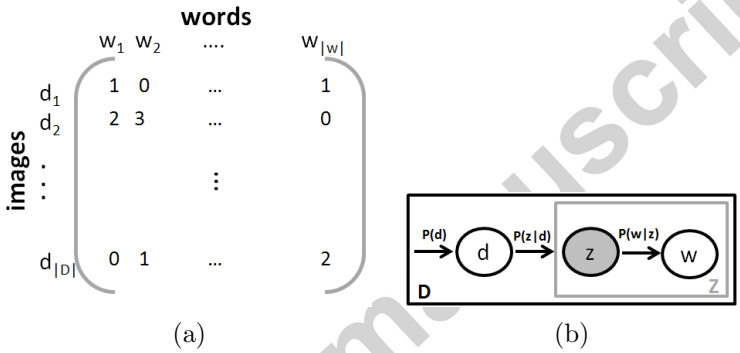

Figure 1: a) co-occurrence data table $n(d, w)$ for images and words, b) the standard pLSA model.

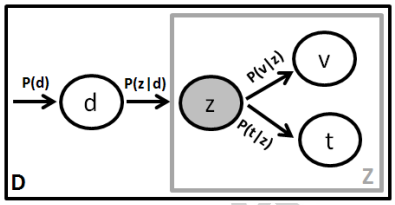

Figure 2: Graphical representation of the *highOrder-plsa* model

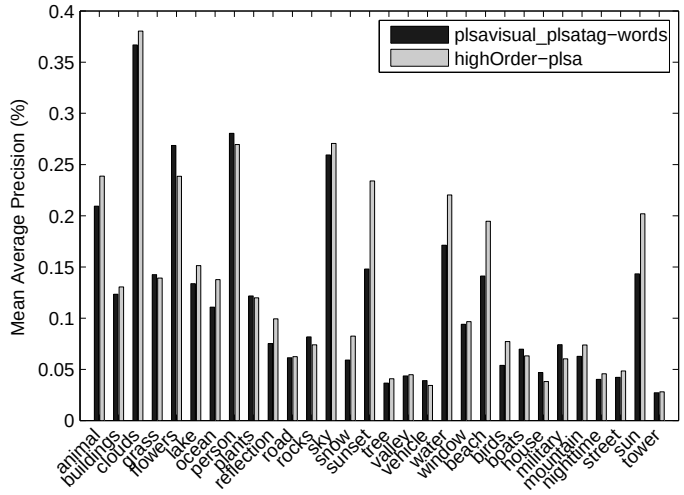

Figure 3: Performance scores on a concept-basis - NUS_WIDE Dataset

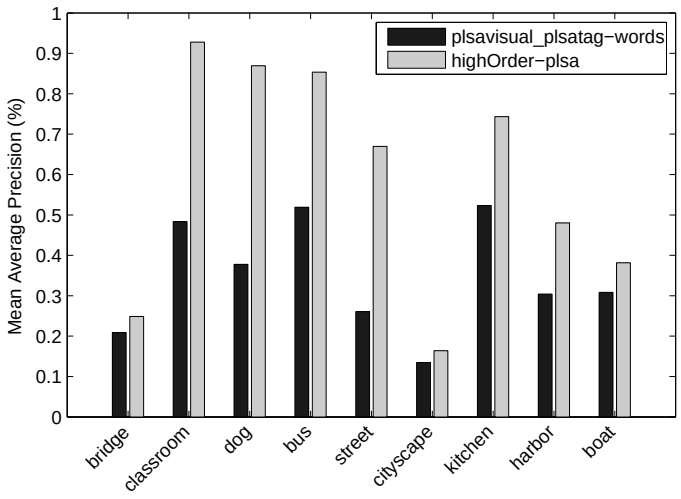

Figure 4: Performance scores on a concept-basis - SOCIAL20 Dataset

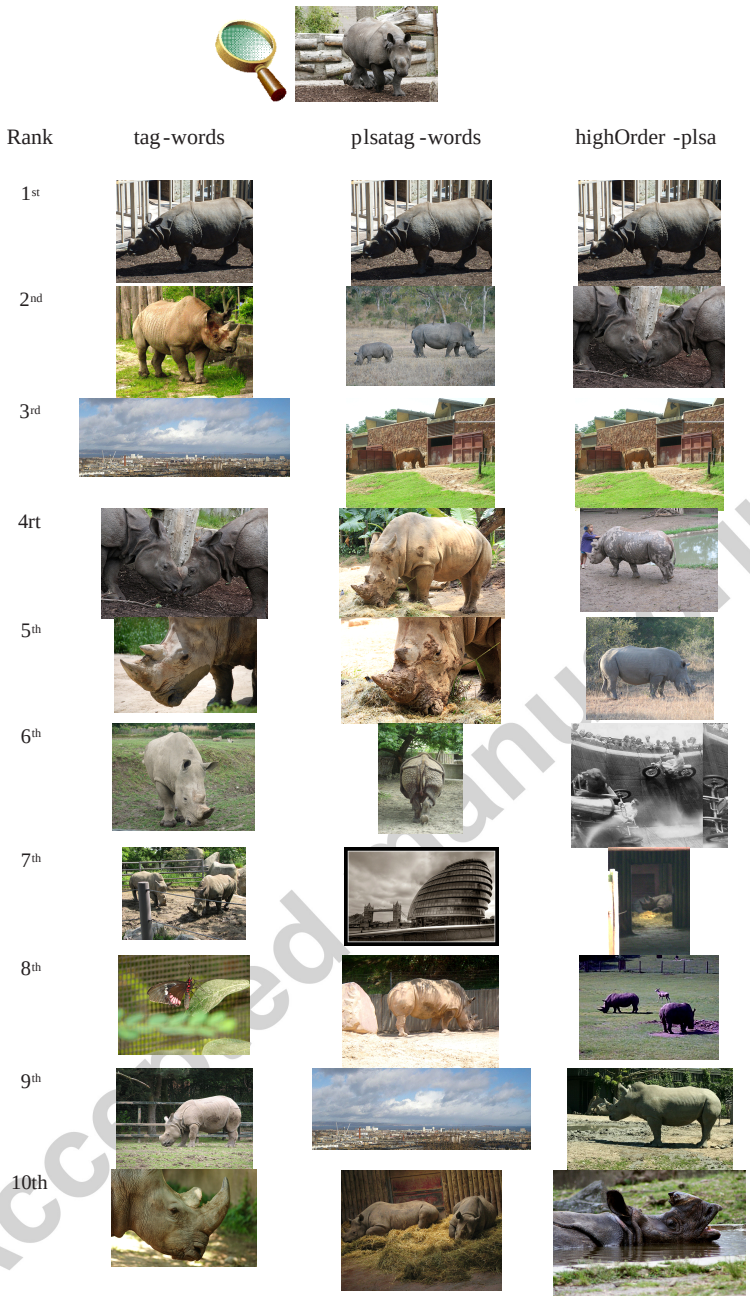

Figure 5: Indicative retrieval examples using the tag-words, plsatag-words and the cross-words-profiles approach for generating the feature space.

Figure 6: Indicative retrieval examples using the visual-words, plsavisual-words and the cross-words-profiles approach for generating the feature space.

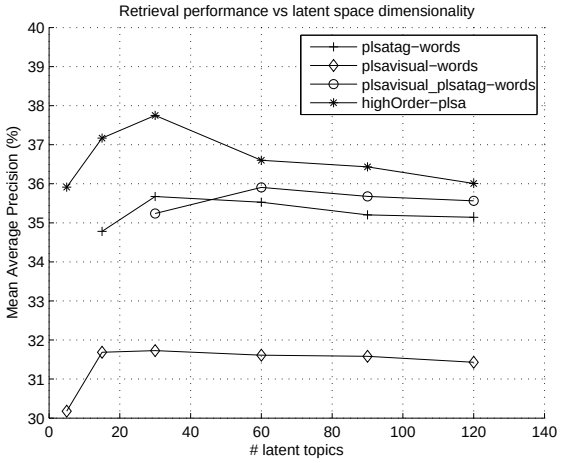

Figure 7: Impact of the latent space dimensionality on the retrieval performance

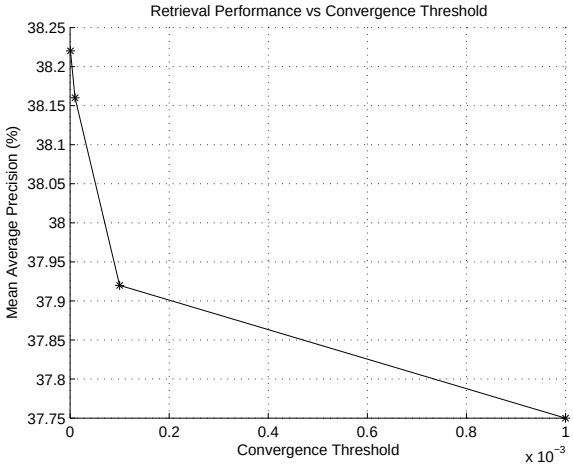

Figure 8: Impact of the convergence threshold on the retrieval performance

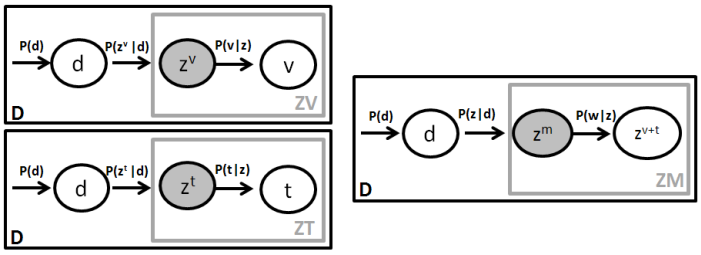

Figure 9: Graphical representation of the *ml-plsa* model [16]

List of Tables

Table 1: Performance scores for image retrieval - Mean Average Precision (%)

Feature Space	**#dims**	**NUS_WIDE**	**SOCIAL20**
tag-words	1000	29,45	26,76
visual-words	500	31,07	10,38
visualtag-words	1500	31,08	10,38
avgnorm-visualtag-words	-	30,22	27,93
cca-visualtag-words	1000	29,88	5,6
plsatag-words	30	35,674	33,92
plsavisual-words	30	31,728	10,69
plsavisual_plsatag-words	60	35,906	34,54
highOrder-plsa	**30**	**37,75**	**59,40**

Table 2: Performance scores for image clustering

Feature Space	**#dims**	**NMI**
tag-words	1000	0.0448
visual-words	500	0.0164
visualtag-words	1500	0.0166
avgnorm-visualtag-words	-	0.0204
cca-visualtag-words	1000	0.0187
plsatag-words	30	0.06591
plsavisual-words	30	0.01977
plsavisual_plsatag-words	60	0.04809
highOrder-plsa	**30**	**0.07979**

Table 3: Execution time for different calculation models

Method	**Calculation model**	**Elapsed Time (sec)**	
		Train	Test
plsatag-words	-	946	78
plsavisual-words	-	250	17
plsavisual-plsatag-words	-	1196	95
highOrder-plsa	Centralized	12288	2502
	Distributed (Memory)	3563	349
	Distributed (Disk)	6687	549

Table 4: Performance scores for image retrieval - Full NUS_WIDE Dataset

Feature Space	**#dims**	**MAP (%)**
tag-words	1000	29,90
visual-words	500	30,470
visualtag-words	1500	30,476
cca-visualtag-words	1000	29.25
plsatag-words	30	35,512
plsavisual-words	30	31,128
plsavisual_plsatag-words	60	35,686
highOrder-plsa	**30**	**38,50**

Table 5: Performance scores for image retrieval - Mean Average Precision (%)

Feature Space	**#dims**	**NUS_WIDE**	**SOCIAL20**
ml-plsa	30	35,956	35,05
mm-plsa	30	34,162	33,36
count-plsa	30	35,266	33,12
highOrder-plsa	**30**	**37,75**	**59,40**

Research Highlights

1. **H1:** We combine the visual and tag information of images.
2. **H2:** We use high order pLSA for generating a semantics sensitive latent space.
3. **H3:** We introduce the concept of word-profiles for measuring the cross-words dependency.
4. **H3:** We implement a distributed calculation model for high order pLSA.
5. **H4:** We evaluate our approach for the tasks of image retrieval and clustering.