Social Image Search exploiting Joint Visual-Textual information within a Fuzzy Hypergraph Framework

Konstantinos Pliakos [#1], Constantine Kotropoulos [#2]

[#] *Department of Informatics, Aristotle University of Thessaloniki*
Thessaloniki, 54124, Greece
[1] kpliakos@aiia.csd.auth.gr
[2] costas@aiia.csd.auth.gr

***Abstract*—The unremitting growth of social media popularity is manifested by the vast volume of images uploaded to the web. Despite the extensive research efforts, there are still open problems in accurate or efficient image search methods. The majority of existing methods, dedicated to image search, treat the image visual content and the semantic information captured by the social image tags, separately or in a sequential manner. Here, a novel and efficient method is proposed, exploiting visual and textual information simultaneously. The joint visual-textual information is captured by a fuzzy hypergraph powered by the term-frequency and inverse-document-frequency (tf-idf) weighting scheme. Experimental results conducted on two datasets substantiate the merits of the proposed method. Indicatively, an average precision of 77% is measured at 1% recall for image-based queries.**

I. Introduction

Nowadays, social image search has evolved into a problem of crucial importance, due to the rising popularity of social media. The development of multimedia and network technologies has led to a vast volume of data uploaded to the web. A great number of social media sharing platforms have been developed, enabling their users to upload images and to annotate them, describing the image content and providing text information. The majority of media search methods rely solely on the text information, achieving unsatisfactory results. The main reason is the noise in user-provided text information. Tags suffer from several well-known limitations, such as ambiguity and lack of uniformity. Many tags are irrelevant, incorrectly spelled, or even completely false. Consequently, new image search methods are certainly needed.

Extensive research efforts have been made in the field of social image search. As a solution to the problem of the noise in tags, L. Chen *et al.* [1] proposed a batch mode re-tagging method for tag refinement. Other, tag refinement methods were proposed in [2] and [3]. In [2], a graphical model, named regularized Latent Dirichlet Allocation (rLDA) was developed in order to jointly estimate both tag similarity and tag relevance. In [3], tag refinement was cast as a tag matrix decomposition into a low rank refined matrix and a sparse error matrix. In [4], a diverse relevance ranking scheme was proposed, taking into account both relevance and diversity. The relevance scores of images were first estimated and the final ranking list was generated by a greedy algorithm, optimizing the average diverse precision. In [5], an image ranking and retrieval approach was elaborated using the correlations between the attributes of a query and the vocabulary terms not present in the query. In [6], a web image search reranking method was proposed, based on a multimodal graph-based learning scheme. Multimodal image feature sets, such as color, edge, and texture were integrated into the aforementioned scheme. The relevance scores, the weights of modalities, and the distance metric for each modality were learned simultaneously.

A hypergraph is a generalization of a graph having edges, which link more than two vertices, that are named as hyperedges. Hypergraphs have been employed in many tasks, such as image retrieval [7]–[9], image classification [10], [11] and object recognition [12], [13]. Recently, hypergraphs have been widely adopted in personalized multimedia recommendation tasks [14]–[17], because of their ability to capture higher-order relationships.

Here, a novel approach to the social image search problem is proposed, using learning on a fuzzy hypergraph powered by term-frequency and inverse-document-frequency (tf-idf) weighting scheme [18]. This way, the fact that some terms are generally more common than others is moderated. Motivated by [19], a hypergraph model was employed, where the images are modeled as vertices and each visual or textual term generates a hyperedge. This way, both visual and textual information is utilized simultaneously, in a joint hypergraph learning framework. Here, in contrast to [19], a fuzzy hypergraph model is employed instead of a binary one and the hyperedge weights are set to 1 to minimize the computational time. In the proposed approach, both image queries and text (tag) queries are taken into account. The experimental results demonstrate the merits of the approach. Indicatively, an average precision of 77% is measured at 1% recall for image-based queries.

The remainder of this paper is organized as follows. In Section II, the ranking on a hypergraph, is detailed. In Section III, the fuzzy hypergraph construction is explained. Experi-

2014 IEEE 16th International Workshop on Multimedia Signal Processing (MMSP), Sep. 22-24, 2014, Jakarta, Indonesia.
978-1-4799-5896-2/14/$31.00

TABLE I
NOTATION SUMMARIZATION

Notation	Definition
κ	The number of visual terms.
λ	The number of text terms.
n	The number of heperedges (i.e. $n = \kappa + \lambda$).
m	The number of images.
$\mathbf{I}$	The identity matrix.
Ψ	The hypergraph.
V	The set of hypergraph vertices. (Here, $\|V\| = m$.)
E	The set of hypergraph hyperedges. (Here, $\|E\| = n$.)
$\mathbf{D}_v$	The $\|V\| \times \|V\|$ diagonal vertex degree matrix.
$\mathbf{D}_e$	The $\|E\| \times \|E\|$ diagonal hyperedge degree matrix.
$\mathbf{W}$	The $\|E\| \times \|E\|$ diagonal matrix containing the hyperedge weights.
$\mathbf{H}$	The $\|V\| \times \|E\|$ incidence matrix for the hypergraph.
$\delta(v)$	The vertex degree.
$\delta(e)$	The hyperedge degree.
$w(e)$	The hyperedge weights.
ϑ	The regularizing parameter.
Ω	The cost function.
$\mathbf{L}$	$\|V\| \times \|V\|$ Zhou's Laplacian matrix of the hypergraph.
$\tilde{g}$	The quantized codevector representation of an image.
$\mathbf{f}$	The ranking vector $\mathbf{f} \in \mathbb{R}^{\|V\|}$.
$\mathbf{y}$	The query vector $\mathbf{y} \in \mathbb{R}^{\|V\|}$.
$G = \{g_1, g_2, \cdots, g_m\}$	The image set, containing m images.
$\Gamma = \{\gamma_1, \gamma_2, \cdots, \gamma_K\}$	The image set, containing K images returned by a simple tag-based search.
$D_G = \{d_{g_1}, d_{g_2}, \cdots, d_{g_m}\}$	The document-image set, containing m image derived documents.
$Z_s = \{z_{s_1}, z_{s_2}, \cdots, z_{s_\kappa}\}$	The visual term set, containing κ visual terms.
$Z_t = \{z_{t_1}, z_{t_2}, \cdots, z_{t_\lambda}\}$	The textual term set, containing λ textual terms.
$\mathbf{B}_t$	The $m \times \lambda$ document-term matrix.
$\mathbf{B}_s$	The $m \times \kappa$ image-visual term matrix.
f_{ij}	The occurrence frequency of term j in image i.

mental results are provided in Section IV, demonstrating the effectiveness of the proposed method. Conclusions are drawn and topics for future research are indicated in Section V.

II. GENERAL HYPERGRAPH MODEL

A hypergraph is a generalization of a graph with edges linking more than two vertices. This way, the hypergraph can capture higher-order information. Hereafter, set cardinality is denoted by $|\cdot|$, the ℓ_2 norm of a vector appears as $\|.\|_2$. All other notation is summarized in Table I. Let $\Psi(V, E, w)$ denote a hypergraph, with set of vertices V and set of hyperedges E to which a weight function $w{:}E \to \mathbb{R}$ is assigned. An incidence matrix $\mathbf{H}$ of size $|V| \times |E|$ has elements:

$$H(v,e) = \begin{cases} 1 & \text{if } v \in e \\ 0 & \text{otherwise.} \end{cases} \tag{1}$$

The vertex and hyperedge degrees are defined as:

$$\left.\begin{aligned} \delta(v) &= \textstyle\sum_{e \in E} w(e) H(v,e) \\ \delta(e) &= \textstyle\sum_{v \in V} H(v,e) \end{aligned}\right\}. \tag{2}$$

Let $\mathbf{A} = \mathbf{D}_v^{-1/2}\mathbf{H}\mathbf{W}\mathbf{D}_e^{-1}\mathbf{H}^T\mathbf{D}_v^{-1/2}$. $\mathbf{A}$ is a symmetric matrix as the diagonal matrices $\mathbf{W}$ and $\mathbf{D}_e{}^{-1}$ commute in multiplication. Then, $\mathbf{L} = \mathbf{I} - \mathbf{A}$ is known as Zhou's normalized Laplacian of the hypergraph [10]. The elements of $\mathbf{A}$, $A(j,i)$, indicate the relatedness between the vertices j and i. To perform clustering on a hypergraph one is seeking for a real-valued ranking vector $\mathbf{f} \in \mathbb{R}^{|V|}$, minimizing the cost function:

$$\Omega(\mathbf{f}) = \mathbf{f}^T\mathbf{L}\mathbf{f}. \tag{3}$$

That is, one requires all vertices with the same value in the ranking vector $\mathbf{f}$ to be strongly connected [20]. $\Omega(\mathbf{f})$ is small, if vertices with high affinities have the same label [20]. For instance, two images are probably similar, if they have many common visual and textual terms (i.e., the hyperedges of the hypergraph). The aforementioned optimization problem was extended to solve a recommendation problem by including the ℓ_2 regularization norm between the ranking vector $\mathbf{f}$ and a query vector $\mathbf{y} \in \mathbb{R}^{|V|}$ [14]. This guarantees that the ranking results do not differ too much from the initial query. The function to be minimized is then expressed as

$$\tilde{Q}(\mathbf{f}) = \Omega(\mathbf{f}) + \vartheta \, ||\mathbf{f} - \mathbf{y}||_2^2 \tag{4}$$

where ϑ is a regularizing parameter. The best ranking vector, $\mathbf{f}^* = \arg\min_{\mathbf{f}} \tilde{Q}(\mathbf{f})$, is found to be [14]:

$$\mathbf{f}^* = \frac{\vartheta}{1+\vartheta}\left(\mathbf{I} - \frac{1}{1+\vartheta}\mathbf{A}\right)^{-1}\mathbf{y}. \tag{5}$$

III. FUZZY HYPERGRAPH CONSTRUCTION

Let each social image $g \in G$ denote a vertex in the hypergraph. Each visual-textual term generates a hyperedge. In Fig. 1, a description of the proposed approach is demonstrated.

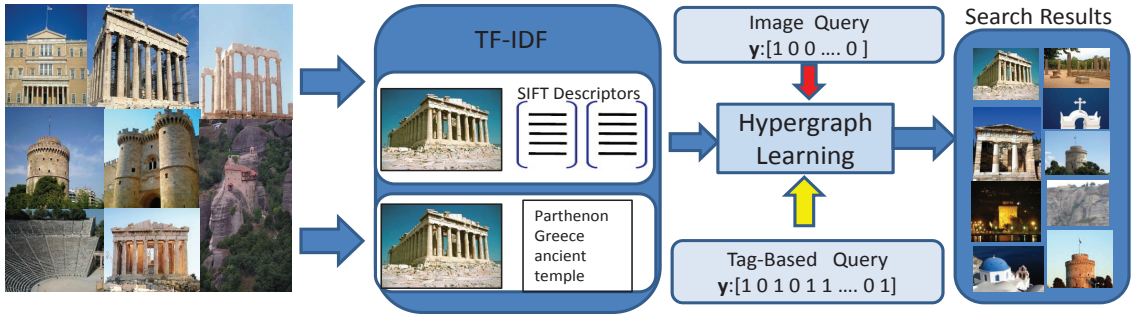

Fig. 1. A description of the proposed social image search approach.

A. Visual Information

Regarding the visual image content, scale-invariant feature transform (SIFT) [21] is employed and SIFT descriptors are extracted from any image g. K-means is applied to the SIFT descriptors of any image g in order to quantize them to a predefined number (e.g., 200) of clusters represented by their mean vectors as codevectors. Accordingly, $g \in G$ is represented by the concatenation of the codevectors $\tilde{g}$, instead of the concatenation of SIFT descriptors. K-means is applied to the set of the aforementioned quantized representations $\tilde{g}$ in order to create the visual word vocabulary. The indices of the resulting codevectors are treated as visual words $z_s \in Z_s = \{z_{s_1}, z_{s_2}, \cdots, z_{s_\kappa}\}$, where κ is the size of visual word vocabulary. Let $\mathbf{B}_s \in \mathbb{R}^{m \times \kappa}$ be the matrix having as elements the term frequency-inverse document frequency (tf-idf) measurements given by $B_s(i,j) = f_{ij} \log_2 \frac{m}{m_j}$, where f_{ij} is the frequency of term j in image i, m_j is the number of images that visual term j appears in, and m is the total number of images.

B. Textual Information

The textual information provided by the tags of social images is captured by following the bag-of-textual-words representation. In order to form a proper text vocabulary, all characters are converted to lower case, unreadable symbols and redundant information are removed. Next, a vocabulary of unique terms is generated along with their frequencies. Then, only the λ terms with the highest frequency are kept. Here, each image set of tags is considered as a document associated to the image. Let $z_t \in Z_t = \{z_{t_1}, z_{t_2}, \cdots, z_{t_\lambda}\}$ be the text vocabulary and $\mathbf{B}_t \in \mathbb{R}^{m \times \lambda}$ be the document-term matrix. Any document-image $d_g \in D_G$ is represented by a vector of size λ, having as elements the tf-idf measurements obtained by taking into account the text information.

C. Hypergraph Learning

Tf-idf is a numerical statistic, quantifying the importance of a term in a document, belonging to a corpus. It is the product of two statistics, namely the term frequency and the inverse document frequency. Term frequency weighs more heavily the terms, which occur often in a specific document. On the other hand, inverse document frequency down-weighs the terms, which tend to appear many times in several documents in the corpus. This way, the fact that some terms are more common than others is handled effectively and thus, the terms that are truly representative of a document are given higher weights.

Here, the binary incidence matrix $\mathbf{H}$ used in [19] is replaced by the fuzzy incidence matrix $\mathbf{H} = \left[\tilde{\mathbf{B}}_s \mid \tilde{\mathbf{B}}_t\right]$ of size $m \times n$, yielding a fuzzy hypergraph model, where $n = \kappa + \lambda$. $\tilde{\mathbf{B}}_s$ and $\tilde{\mathbf{B}}_t$ matrices are obtained by a min-max normalization of $\mathbf{B}_s$ and $\mathbf{B}_t$, respectively, so that their elements admit values in $[0, 1]$. As can be seen, the incidence matrix captures both the visual and the textual information, which is also inherited by the Laplacian of the hypergraph and the matrix $\mathbf{A}$, appearing in (5). To assess the impact of the fuzzy hypergraph incidence matrix the diagonal matrix $\mathbf{W}$ containing the hyperedge weights is set to $\mathbf{I}$ i.e., $w(e) = 1, \forall e \in E$.

In the proposed approach, the query is considered based on either text (tags) or images. In the case of an image-based query, the query vector $\mathbf{y}$ is initialized by setting the entry corresponding to the query image to 1. In the case of a tag-based query, a simple tag-based search method is employed and the K top images are returned from all the images that include the query tag in their corresponding set of tags. Let $\Gamma = \{\gamma_1, \gamma_2 \cdots, \gamma_K\} \subset G$ be the image set associated to this search. The query vector $\mathbf{y}$ is initialized as follows:

$$y(g) = \begin{cases} 1, & \text{if } g \in \Gamma \\ 0, & \text{otherwise.} \end{cases} \tag{6}$$

The ranking vector $\mathbf{f}^*$ is derived by solving (5), as detailed in Section II. It has the same size and structure as $\mathbf{y}$.

IV. Experimental Results

The averaged Recall-Precision, the Mean Average Precision MAP, and F_1 measure are used as figures of merit. Precision is defined as the number of correctly retrieved images divided by the number of all retrieved images. Recall is defined as the number of correctly retrieved images divided by the number of all images. The F_1 measure is the weighted harmonic mean of precision and recall, which measures the effectiveness of retrieval when treating precision and recall as equally

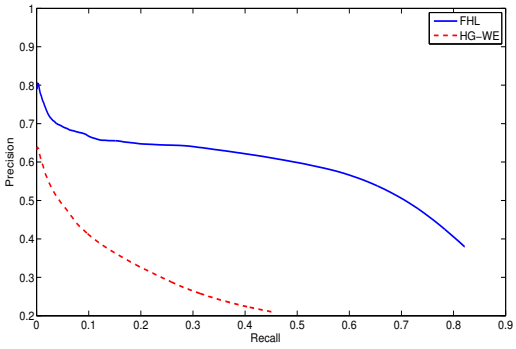

Fig. 2. Averaged Recall-Precision curves for image-based queries.

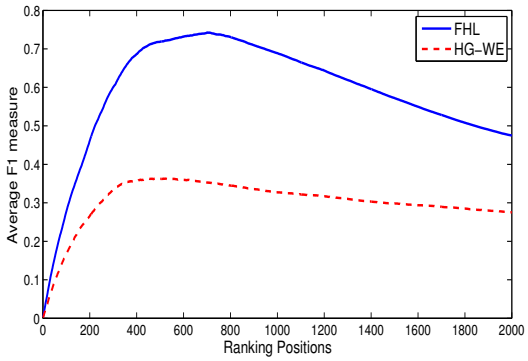

Fig. 3. Averaged F_1 measure at several ranking positions for tag-based queries.

important, as is shown below.

$$F_1 = 2\frac{Precision \cdot Recall}{Precision + Recall} \tag{7}$$

The MAP is the mean value of the Average Precision AP of all the queries. The AP is defined as the average of precisions computed at the point of each correctly retrieved item, as is shown below:

$$AP = \frac{\sum_{i=1}^{Num} Precision@i \cdot true_i}{cNum} \tag{8}$$

where $Precision@i$ is the precision at ranking position i, Num the number of retrieved items, $cNum$ the number of correctly retrieved items, and $true_i = 1$ if the item at position i is correctly retrieved. Let us denote the proposed method as fuzzy hypergraph learning (FHL) and the one proposed in [19] as HG-WE. The HG-WE was implemented by following precisely the details in [19].

In order to evaluate the proposed method, a dataset of 3291 images, depicting 11 popular Greek sites (old city of Rhodes, Santorini, White Tower of Thessaloniki, Parthenon, Delphi, Meteora, ancient Olympia, Sounio, Mycenae, Greek Parliament, and Epidaurus) was collected from $Flickr$[1]. These Greek sites were used as query tags in the evaluation procedure. Both textual and visual vocabularies were derived and the typical values of κ and λ were set to 3000 and 2000, respectively. Next, the dataset was further enriched by including 4986 unseen test images collected also from $Flickr$. Experiments were conducted for both image-based queries ($K = 1$) and tag-based queries ($K = 100$).

The FHL clearly outperforms the HG-WE for both image-based and tag-based search, as shown in Figures 2 and 3, respectively. As is demonstrated in Figure 2, a precision rate of 77% is achieved for 1% recall for image-based queries. For tag-based queries, the maximum average F_1 measure equals 0.7425. It is obtained at the ranking position 704. The complete curve of the averaged F_1 measure per ranking position is displayed in Figure 3.

For further evaluation, the NUS-WIDE dataset [22] was employed. A subset of 39450 images corresponding to 10 tags (leaf, grass, swimmers, coral, mountain, soccer, map, running, sun, sunset) was extracted for further processing. In Table II, the averaged F_1 measure corresponding to 4 different ranking positions and the MAP are listed for the FHL and the HG-WE, for tag-based queries. It is clearly seen that the FHL outperforms the HG-WE.

TABLE II
F_1 MEASURE AND MAP VALUES FOR FHL AND HG-WE.

	F_1@20	F_1@200	F_1@1000	F_1@2000	MAP
FHL	0.1452	0.3232	0.3393	0.3116	0.4695
HG-WE	0.1125	0.1592	0.13874	0.1374	0.2341

At this point, it has to be mentioned that the FHL is much less computationally expensive than the HG-WE, as only one least squares minimization problem should be solved to obtain $\mathbf{f}^*$ in (5).

V. Conclusion

Here, an efficient social image search approach was proposed. It is based on a fuzzy hypergraph learning framework generated by tf-idf weighting. The noise in user-provided text information is handled by using the tf-idf weighting scheme to jointly model the visual and textual information. The proposed method outperforms the existing hypergraph-based learning methods. The experimental results demonstrated the efficiency of the proposed method for both tag-based and image-based queries that do not require significant computational effort.

In the future, the proposed approach will be enhanced by integrating more effective visual features within the framework and by applying more efficient iterative optimization solutions to the hypergraph learning algorithm.

Acknowledgment

This research has been co-financed by the European Union (European Social Fund - ESF) and Greek national funds through the Operation Program "Competitiveness-Cooperation 2011" - Research Funding Program: 11SYN-10-1730-ATLAS.

[1] http://www.flickr.com

REFERENCES

[1] L. Chen, D. Xu, I. W. Tsang, and J. Luo, "Tag-based web photo retrieval improved by batch mode re-tagging," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, June 2010, pp. 3440–3446.

[2] H. Xu, J. Wang, X. S. Hua, and S. Li, "Tag refinement by regularized lda," in *Proc. ACM Int. Conf. Multimedia*, 2009, pp. 573–576.

[3] G. Zhu, S. Yan, and Y. Ma, "Image tag refinement towards low-rank, content-tag prior and error sparsity," in *Proc. ACM Int. Conf. Multimedia*, 2010, pp. 461–470.

[4] M. Wang, K. Yang, X. S. Hua, and H. J. Zhang, "Towards a relevant and diverse search of social images," *IEEE Trans. Multimedia*, vol. 12, no. 8, pp. 829–842, 2010.

[5] B. Siddiquie, R. S. Feris, and L. S. Davis, "Image ranking and retrieval based on multi-attribute queries," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, June 2011, pp. 801–808.

[6] M. Wang, H. Li, D. Tao, K. Lu, and X. Wu, "Multimodal graph-based reranking for web image search," *IEEE Trans. Image Processing*, vol. 21, no. 11, pp. 4649–4661, 2012.

[7] Q. Liu, Y. Huang, and D. N. Metaxas, "Hypergraph with sampling for image retrieval," *Pattern Recognition*, vol. 44, no. 10, pp. 2255–2262, 2011.

[8] Y. Huang, Q. Liu, S. Zhang, and D. Metaxas, "Image retrieval via probabilistic hypergraph ranking," in *Proc. IEEE Int. Conf. Computer Vision Pattern Recognition*, 2010, pp. 3376–3383.

[9] Y. Liu, J. Shao, J. Xiao, F. Wu, and Y. Zhuang, "Hypergraph spectral hashing for image retrieval with heterogeneous social contexts," *Neurocomputing*, vol. 119, pp. 49–58, 2013.

[10] D. Zhou, J. Huang, and B. Schölkopf, "Learning with hypergraphs: Clustering, classification, and embedding," in *Advances in Neural Information Processing Systems*, 2006, pp. 1601–1608.

[11] J. Yu, D. Tao, and M. Wang, "Adaptive hypergraph learning and its application in image classification," *IEEE Trans. Image Processing*, vol. 21, no. 7, pp. 3262–3272, 2012.

[12] S. Xia and E. R. Hancock, "3d object recognition using hyper-graphs and ranked local invariant features," in *Structural, Syntactic, and Statistical Pattern Recognition*, 2008, pp. 117–126.

[13] Y. Gao, M. Wang, D. Tao, R. Ji, and Q. Dai, "3-d object retrieval and recognition with hypergraph analysis," *IEEE Trans. Image Processing*, vol. 21, no. 9, pp. 4290–4303, 2012.

[14] J. Bu, S. Tan, C. Chen, C. Wang, H. Wu, Z. Lijun, and X. He, "Music recommendation by unified hypergraph: Combining social media information and music content," in *Proc. ACM Conf. Multimedia*, 2010, pp. 391–400.

[15] J. Xu, V. Singh, Z. Guan, and B. S. Manjunath, "Unified hypergraph for image ranking in a multimodal context," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, 2012, pp. 2333–2336.

[16] K. Pliakos and C. Kotropoulos, "Simultaneous image tagging and geo-location prediction within hypergraph ranking framework," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, May 2014, pp. 6944–6948.

[17] ——, "Personalized and geo-referenced image recommendation using unified hypergraph learning and group sparsity optimization," in *Proc. IEEE Int. Symp. Communications, Control, and Signal Processing*, May 2014, pp. 323–326.

[18] G. Salton, A. Wong, and C. Yang, "A vector space model for automatic indexing," *Communications of the ACM*, vol. 18, no. 11, pp. 613–620, 1975.

[19] Y. Gao, M. Wang, Z. J. Zha, J. Shen, X. Li, and X. Wu, "Visual-textual joint relevance learning for tag-based social image search," *IEEE Trans. Image Processing*, vol. 22, no. 1, pp. 363–376, 2013.

[20] S. Agarwal, K. Branson, and S. Belongie, "Higher order learning with graphs," in *Proc. 23rd Int. Conf. Machine Learning*, 2006, pp. 17–24.

[21] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. Journal Computer Vision*, vol. 60, no. 2, pp. 91–110, 2004.

[22] T. S. Chua, J. Tang, R. Hong, H. Li, Z. Luo, and Y. Zheng, "Nus-wide: a real-world web image database from national university of singapore," in *Proc. ACM Int. Conf. Image Video Retrieval*, 2009, pp. 1–9.