DHCN: Deep Hierarchical Context Networks For Image Annotation

Mingyuan Jiu, Hichem Sahbi

DHCN: DEEP HIERARCHICAL CONTEXT NETWORKS FOR IMAGE ANNOTATION

Mingyuan Jiu[⋆,†] *Hichem Sahbi*[‡]

[⋆] School of Information Engineering, Zhengzhou University, Zhengzhou, China
[†] Zhengzhou University Research Institute of Industrial Technology Co., Ltd., Zhengzhou, China
[‡] Sorbonne University, UPMC, CNRS, LIP6, F-75005, Paris, France

ABSTRACT

Context modeling is one of the most fertile sub-fields of visual recognition which aims at designing discriminant image representations while incorporating their intrinsic and extrinsic relationships. However, the potential of context modeling is currently under-explored and most of the existing solutions are either context-free or restricted to simple handcrafted geometric relationships.

We introduce in this paper DHCN: a novel Deep Hierarchical Context Network that leverages different sources of contexts including geometric and semantic relationships. The proposed method is based on the minimization of an objective function mixing a fidelity term, a context criterion and a regularizer. The solution of this objective function defines the architecture of a bi-level hierarchical context network; the first level of this network captures scene geometry while the second one corresponds to semantic relationships. We solve this representation learning problem by training its underlying deep network whose parameters correspond to the most influencing bi-level contextual relationships and we evaluate its performances on image annotation using the challenging ImageCLEF benchmark.

Index Terms— Hierarchical context learning, deep context-aware networks, image annotation.

1. INTRODUCTION

Image annotation is one of the major challenges in computer vision which aims at assigning keywords (a.k.a labels or concepts) to images. The difficulty in image annotation stems from the extreme variability of the learned concepts and their versatile content which is usually described with handcrafted or learned representations [1–11]. However, due to its limited representational power, content is usually upgraded with context in order to capture both the intrinsic and the extrinsic properties of images[1]. Indeed, while context-free models are effective when images (from the same concepts) are well clustered, they miserably fail when concepts exhibit a strong intra-class variability. In contrast, context-dependent solutions reduce the ratio between intra and inter class variability even when content of images — belonging to the same concepts — is corrupted [12].

Several existing methods leverage context prior to achieve image annotation and outperform context-free approaches by a significant margin. In these solutions, context is usually defined as a neighborhood system, i.e., a set of geometric or statistical dependencies between low level primitives (such as interest points, regions, etc.) or semantic relationships. These relationships make it possible to model pairwise and high-order interactions between images and their primitives using well designed objective functions; several works follow this line including neighborhood embedding [13] and spatially-constrained deep learning [14, 15]. These methods learn functions that map neighboring data from the input (raw) space into a well designed feature space while maintaining their proximity. Other methods rely on structural regularization which integrates a priori knowledges into different penalization terms and constrain the learned models to reflect these knowledges. Typical works include ℓ_1-norm [16], ℓ_0-norm, ℓ_{12}-norm [17] and structural regularization [18] which usually define convex (globally optimal) problems. Variants of these models consider prediction scores on labeled and unlabeled data for regularization (as in Laplacian SVMs [19, 20]) in order to diffuse labels from training to test data. More recently, graph neural networks have attracted a particular attention as an extension of convolutional neural networks (CNNs) [21–25] to non-Euclidean domains [26–31] and have shown very promising performances on relational graph data.

The success of all the aforementioned methods is very dependent on the relevance of the used neighborhood systems which are usually handcrafted and when learned they are restricted only to simple geometric relationships. In contrast, the solution proposed in this paper learns both geometric and semantic contextual relationships in a unified framework. Our

This work was supported by a grant from the National Natural Science Foundation of China (No. 61806180, U1804152), by a grant from Key Research Projects of Henan Higher Education Institutions in China (No. 19A520037), by a grant from Science and Technology Innovation Project of Zhengzhou (2019CXZX0037), and also in part by a grant from the research agency ANR (Agence Nationale de la Recherche) of France under the MLVIS project (ANR-11-BS02-0017).

[1] Intrinsic properties of images are usually related to scene structure or geometry while extrinsic properties refer to semantic relationships (such as "image-to-image" links in social networks).

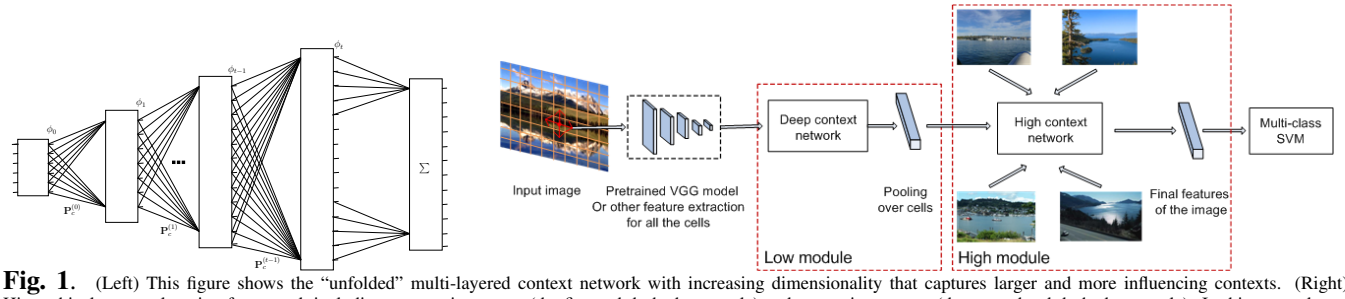

Fig. 1. (Left) This figure shows the "unfolded" multi-layered context network with increasing dimensionality that captures larger and more influencing contexts. (Right) Hierarchical context learning framework including geometric context (the first red dashed rectangle) and semantic contexts (the second red dashed rectangle). In this network, we only show four semantic neighbors for each image as an example **(better to zoom the PDF version)**.

design principle relies on the context-dependent similarities introduced in [12, 32–34] but, in contrast to these works, considers learned bi-level contexts instead of handcrafted ones. Learning context translates into optimizing the adjacency matrices of the neighborhood system and this equivalently reduces to training a particular deep network whose parameters correspond to the most influencing (geometric and semantic) relationships in image annotation. With this approach, the representation of a given image is obtained recursively by aggregating (i) the representations of neighboring primitives (insides images) following the learned geometric context and (ii) those of neighboring images according to the learned semantic context. This results into a highly discriminant hierarchical representation as shown later in experiments.

2. CONTEXT-AWARE SIMILARITY NETWORKS

Let $\mathcal{I} = \{\mathcal{I}_p\}_{p=1}^P$ denote a collection of training images and $\mathcal{S}_p = \{\mathbf{x}_1^p, \ldots, \mathbf{x}_n^p\}$ be a list of non-overlapping cells taken from a regular grid of $\mathcal{I}_p$; without a loss of generality, we assume n constant for all images. A context-aware similarity (or kernel denoted as κ) is a symmetric and positive semi-definite (p.s.d) function that returns the resemblance between any two given cells $\mathbf{x}$, $\mathbf{x}'$ in $\mathcal{X} = \cup_p \mathcal{S}_p$. As designed subsequently, the particularity of κ w.r.t. many usual kernels (such as linear, RBF, etc.) is that $\kappa(\mathbf{x}, \mathbf{x}')$ depends *not only* on the content of the cells $(\mathbf{x}, \mathbf{x}')$ *but also* on their context $\{\mathcal{N}_c(\mathbf{x}) \times \mathcal{N}_c(\mathbf{x}')\}_c$; here $\{\mathcal{N}_c(\mathbf{x})\}_c$ corresponds to the neighborhood system, i.e., the set of neighbors of $\mathbf{x}$ with particular (learned) geometric relationships. The kernel κ (or equivalently its gram matrix $\mathbf{K}$) is learned by minimizing the following objective function

$$\min_{\mathbf{K}} \mathrm{tr}(-\mathbf{K}\mathbf{S}') - \alpha_1 \sum_{c=1}^{C} \mathrm{tr}(\mathbf{K}\mathbf{P}_c\mathbf{K}'\mathbf{P}_c') + \frac{\beta_1}{2}||\mathbf{K}||_2^2, \quad (1)$$

here $'$ and tr denotes matrix transpose and the trace operator respectively, $\kappa(\mathbf{x}, \mathbf{x}') = \mathbf{K}_{\mathbf{x},\mathbf{x}'}$ (with $\mathbf{K}_{\mathbf{x},\mathbf{x}'}$ being an entry of $\mathbf{K}$), $\mathbf{S}$ is a (context-free) visual similarity matrix between data in $\mathcal{X}$ and $\alpha_1 \geq 0$, $\beta_1 > 0$ balance similarity between neighboring cells and regularization. In the above objective function the matrices $\{\mathbf{P}_c\}_c$ correspond to a neighborhood system $\{\mathcal{N}_c(.)\}_{c=1}^C$; each entry $\mathbf{P}_{c,\mathbf{x},\mathbf{x}'} \neq 0$ if $\mathbf{x}' \in \mathcal{N}_c(\mathbf{x})$, otherwise $\mathbf{P}_{c,\mathbf{x},\mathbf{x}'} \leftarrow 0$. In practice, C (with $C = 4$) different types of neighbors are considered (top, bottom, left, right) and the initial spatial support of these neighbors $\{\mathcal{N}_c(\mathbf{x})\}_{c=1}^C$ corresponds to a disk with a radius r around $\mathbf{x}$ (see more details about the setting of r in experiments). Using κ, one may define the similarity between any two given images $\mathcal{I}_p$ and $\mathcal{I}_q$ using convolution which aggregates the similarities between all the pairs in $\mathcal{S}_p \times \mathcal{S}_q$ as $\mathcal{K}(\mathcal{S}_p, \mathcal{S}_q) = \sum_{i,j} \kappa(\mathbf{x}_i^p, \mathbf{x}_j^q)$. Note that $\mathcal{K}$ is also symmetric and p.s.d resulting from the closure of the positive semi-definiteness w.r.t. the sum.

One may show that the solution of Eq. (1) is recursively obtain as the fixed-point (denoted as $\tilde{\mathbf{K}}$) of

$$\mathbf{K}^{(t+1)} = \mathbf{S} + \gamma_1 \sum_{c=1}^{C} \mathbf{P}_c \mathbf{K}^{(t)} \mathbf{P}_c', \quad (2)$$

with $\gamma_1 = \alpha_1/\beta_1$. Resulting from the p.s.d of $\{\mathbf{K}^{(t)}\}_t$ (thereby $\tilde{\mathbf{K}}$) and $\mathcal{K}$, the maps associated to these kernels are respectively [33]

$$\begin{aligned} \mathbf{\Phi}^{(t+1)} &= \left(\mathbf{\Phi}'^{(0)} \;\; \gamma_1^{\frac{1}{2}} \mathbf{P}_1 \mathbf{\Phi}'^{(t)} \;\; \ldots \;\; \gamma_1^{\frac{1}{2}} \mathbf{P}_C \mathbf{\Phi}'^{(t)}\right)' \\ \phi_{\mathcal{K}}(\mathcal{S}_p) &= \sum_{\mathbf{x} \in \mathcal{S}_p} \tilde{\mathbf{\Phi}}_{\mathbf{x}}, \end{aligned} \quad (3)$$

here $\tilde{\mathbf{\Phi}}_{\mathbf{x}}$ denotes the restriction of $\tilde{\mathbf{\Phi}}$ to $\mathbf{x}$ and $\mathbf{\Phi}^{(0)}$ is the map of the initial kernel $\mathbf{K}^{(0)}$; for instance, this initial map can be exactly set using the Kronecker tensor product for the polynomial kernel or approximated using KPCA for any other kernel (see more details in [33, 35]). Following the recursive form in Eq. (3), it is easy to see that the latter is strictly equivalent to a multi-layered deep network (also referred to as *deep context network*) whose input is $\mathbf{\Phi}^{(0)}$, intermediate layers $\{\mathbf{\Phi}^{(t)}\}_t$, output $\phi_{\mathcal{K}}(\mathcal{S}_p)$ and weights corresponding to the adjacency matrices $\{\mathbf{P}_c\}_c$ (see Fig. 1, left); hence training this network makes it possible to learn the neighborhood system, i.e., the spatial (geometric) context.

3. DEEP HIERARCHICAL CONTEXT LEARNING

In this section, we extend the previous framework to build a deep hierarchical context network that learns not only geo-

metric but also semantic relationships between images. This turns out to be more effective as shown later in experiments.

3.1. Bi-level context learning

As describe earlier, context learning makes it possible to capture spatial relationships between image cells. While being already performant, this design focuses mainly on the geometric structure of images and ignores totally other types of relationships, namely semantic ones. The tenet in this extension is to consider an extra-level in context-aware similarity design that considers images similar not only when their learned representations $\{\phi_{\mathcal{K}}(\mathcal{S}_p)\}_p$ are close but also when their semantic context is similar too. The notion of semantic context is inherently different but complementary w.r.t. the one used earlier; indeed, the semantic neighborhood system (now denoted as $\mathcal{N}_{\mathcal{I}}(\mathcal{S}_p)$), associated to any given image $\mathcal{S}_p$, is defined as the set of images sharing semantic relations[2] with $\mathcal{S}_p$. Considering $\mathbf{P}_{\mathcal{I}}$ as the adjacency matrix related to $\mathcal{N}_{\mathcal{I}}(.)$, and $\mathbf{K}_{\mathcal{I}}$ the targeted context-aware similarity (to learn), we find the latter by minimizing a variant of Eq. (1)

$$\min_{\mathbf{K}_{\mathcal{I}}} \text{tr}(-\mathbf{K}_{\mathcal{I}}\tilde{\mathbf{S}}') - \alpha_2 \text{tr}(\mathbf{K}_{\mathcal{I}}\mathbf{P}_{\mathcal{I}}\mathbf{K}_{\mathcal{I}}^{'}\mathbf{P}_{\mathcal{I}}^{'}) + \frac{\beta_2}{2}||\mathbf{K}_{\mathcal{I}}||_2^2, \quad (4)$$

here $\alpha_2 \geq 0$, $\beta_2 > 0$, $\mathbf{K}_{\mathcal{I}}$ is the learned similarity matrix for images in $\mathcal{I}$ and entries of $\tilde{\mathbf{S}}$ correspond to inner products of the obtained $\{\phi_{\mathcal{K}}(\mathcal{S}_p)\}_p$ on the fixed-points of Eq. (3). Similarly, one may show that the solution of Eq. (4) can be recursively defined as $\mathbf{K}_{\mathcal{I}}^{(t+1)} = \tilde{\mathbf{S}} + \gamma_2 \mathbf{P}_{\mathcal{I}}\mathbf{K}_{\mathcal{I}}^{(t)}\mathbf{P}_{\mathcal{I}}^{'}$ which is again a p.s.d kernel whose map is explicitly given by

$$\mathbf{\Phi}_{\mathcal{I}}^{(t+1)} = \left(\phi_{\mathcal{K}}'(\mathcal{I}) \;\; \gamma_2^{\frac{1}{2}}\mathbf{P}_{\mathcal{I}}\mathbf{\Phi}_{\mathcal{I}}^{'(t)}\right)', \quad (5)$$

with $\gamma_2 = \alpha_2/\beta_2$ and $\phi_{\mathcal{K}}'(\mathcal{I})$ dependent on the geometric context $\mathbf{P}_c$. By combining the recursive forms in Eqs. (3) and (5), one may define a deep context network (related to Eq. 5) on top of another one (related to Eq. 3); training the parameters $\{\mathbf{P}_c\}_c, \mathbf{P}_{\mathcal{I}}$ of this complete deep hierarchical context network (DHCN) makes it possible to learn bi-level contextual relationships where the first level captures low-order geometric relationships while the second level models high-order semantic links between images. The whole architecture is shown in Fig. (1, right).

3.2. Optimization

The two objective functions shown earlier define the complete architecture of the DHCN but training its parameters (and hence the context) requires another (supervised) loss. Considering a K-label classification task, a multi-class SVM layer (whose parameters denoted as $\{w_k\}_k$) is stacked on top of DHCN for label prediction. Let $\{(\mathcal{I}_p, \mathbf{Y}_k^p)\}_p$ denote the training set of images and their labels with $\mathbf{Y}_k^p = +1$ iff $\mathcal{I}_p$ belongs to class k and $\mathbf{Y}_k^p = -1$ otherwise. The supervised loss used to train our context matrices and SVM parameters is defined as

$$\min_{\{\mathbf{P}_c\}_c, \mathbf{P}_{\mathcal{I}}, w_k} \sum_{k=1}^{K} \frac{1}{2}||w_k||^2 + C_k \sum_{p=1}^{P} \max(0, 1 - \mathbf{Y}_k^p w_k' \mathbf{\Phi}_{\mathcal{I}}(\mathcal{I}_p)). \quad (6)$$

We solve this problem using alternating optimization. First, we fix $\{\mathbf{P}_c\}_c$ and $\mathbf{P}_{\mathcal{I}}$ and optimize the binary SVMs $\{f_k(.) = w_k'\mathbf{\Phi}_{\mathcal{I}}(.)\}_{k=1}^K$ using LIBSVM [36]. Then, we fix the learned SVMs and update the context parameters by gradient descent. Let E denote the loss in Eq. (6), the gradient of E w.r.t. the final kernel map $\mathbf{\Phi}_{\mathcal{I}}(\mathcal{I}_p)$ is given by

$$\frac{\partial E}{\partial \mathbf{\Phi}_{\mathcal{I}}(\mathcal{I}_p)} = -\sum_{p=1}^{P}\sum_{k=1}^{K} C_k \mathbf{Y}_k^p w_k \mathbb{1}_{\{1 - \mathbf{Y}_k^p w_k' \mathbf{\Phi}_{\mathcal{I}}(\mathcal{I}_p))\}}. \quad (7)$$

Using the chain rule [21], we backpropagate this gradient to the previous layers in order to obtain all the gradients of E w.r.t. $\mathbf{P}_{\mathcal{I}}^{(t)}$ and $\{\mathbf{P}_c^{(t)}\}_c$ for $t = T-1, \ldots, 0$. Finally, we update the context matrices using gradient descent. These two iterative steps are repeated till convergence which is observed (in practice) in less than 100 iterations.

4. EXPERIMENTS

In this section, we apply the proposed DHCN to image annotation using the challenging ImageCLEF benchmark. The goal is to predict a list of keywords that best describes the visual content of images. This benchmark includes more than 250k images belonging to 95 concepts (also referred to as keywords); note that the latter are not exclusive, so one may assign multiple keywords to a given image when the scores of the underlying SVMs are positive. As the ground-truth has been released only on the dev set (of 1,000 images), we randomly split this set into two equally-sized subsets, one for training and another for evaluation.

Each image in ImageCLEF is rescaled to a median dimension of 400×500 pixels and partitioned into a regular grid of 8×10 cells. Two types of features are used to describe the contents of the cells: i) Bag-of-Words (BoW) histogram with a SIFT code-book of 500 dimensions and ii) Deep VGG features pretrained on ImageNet ("imagenet-vgg-m-1024") [37]. This VGG-net is composed of five convolutional and three fully-connected layers and the output of the second fully-connected layer is used to describe the content of the cells in the regular grids. The performances are measured using the average F-scores (harmonic means of recall and precision) both at the concept and the sample levels (denoted respectively as MF-C and MF-S) as well as the mean Average Precision (mAP).

In these experiments, we consider a six layer DHCN architecture corresponding to 2 (geometric context) + 1 (pooling)

[2]For instance, one may consider these relations using similarity or links in social networks.

Method	r	$\lvert\mathcal{N}_\mathcal{I}\rvert$	BoW features		VGG-CNN features	
			Lin kernel map	HI kernel map	Lin kernel map	HI kernel map
CF (Context-free)	-	-	39.7/24.4/46.6	41.3/25.1/49.5	45.3/30.8/56.4	45.5/30.1/57.9
DFCN (Deep fixed context network [33])	1	-	40.6/24.6/48.3	42.6/26.3/50.5	45.8/31.2/57.6	46.4/30.7/58.5
DLCN (Deep learned context network [34])	1	-	42.7/26.4/50.5	45.2/26.4/53.9	47.5/32.7/58.7	48.8/32.7/59.9
DHCN (proposed)	1	10	54.6/43.2/64.8	**55.5**/43.4/65.3	56.0/**44.8**/65.6	55.7/**44.7**/65.8
DFCN (Deep fixed context network [33])	5	-	41.0/25.3/48.9	42.9/26.7/51.3	46.8/31.8/57.9	46.9/31.1/58.7
DLCN (Deep learned context network [34])	5	-	44.0/26.6/52.0	45.6/26.2/54.0	47.9/33.2/58.8	48.4/32.7/59.5
DHCN (proposed)	5	10	54.6/39.8/**64.9**	**55.5**/42.0/65.7	56.1/44.0/65.7	**56.5**/43.8/**66.6**
DHCN (proposed)	1	15	**54.7**/**43.6**/64.4	54.8/**43.6**/**66.0**	**56.2**/44.7/**66.3**	56.0/44.4/66.1

Table 1. The performance (in %) of different methods in the test set of ImageCLEF. A triple $\cdot/\cdot/\cdot$ stands for MF-S/MF-C/mAP. In these experiments r corresponds to the radius of the disk that supports geometric context while $\lvert\mathcal{N}_\mathcal{I}\rvert$ corresponds to the size of semantic context.

Fig. 2. Examples of annotation results using context-free representations ("CF"), deep context networks with fixed and learned contexts (resp. denoted as "DFCN" and "DLCN"), as well as deep hierarchical context network ("DHCN"). "GT" refers to ground-truth annotation while the stars mean the presence of a given concept in the test image.

Kernel	MF-S	MF-C	mAP
GMKL [38]	41.3	24.3	49.1
2LMKL [39]	45.0	25.8	54.0
LDMKL [4]	47.8	30.0	58.6
DLCN [34]	48.8	32.7	59.9
DHCN (proposed)	**56.5**	**43.8**	**66.6**

Table 2. Performance comparison w.r.t. the most closely related work.

layers followed by 2 (semantic context) + 1 (SVM) layers. Linear and histogram intersection (HI) maps are used as inputs to the DHCN and two settings of r (the radius of the disk supporting the geometric context) are considered ($r = 1$ and $r = 5$). Note that the initial matrices $\{\mathbf{P}_c\}_c$ and $\mathbf{P}_\mathcal{I}$ (weights of DHCN) are normalized to be row-stochastic, γ_1 and γ_2 are initially set to 1, and $\lvert\mathcal{N}_\mathcal{I}(.)\rvert \in \{10, 15\}$ in practice.

Tab. 1 shows the performances of context-free networks (related to linear and HI kernel maps) vs. deep context networks with three settings: i) matrices in $\{\mathbf{P}_c\}_c$ are handcrafted (DFCN) ii) only $\{\mathbf{P}_c\}_c$ are learned (DLCN) and iii) both $\{\mathbf{P}_c\}_c$, $\mathbf{P}_\mathcal{I}$ are learned (DHCN). From all these results, we observe that the DHCN outperforms all the other settings by a large margin (for different features and kernel map initializations) compared to context-free and handcrafted deep context networks as well as learned ones (where only geometric context is learned); globally, a more influencing impact on performances is observed with neighborhood systems learned with larger values of r and $\lvert\mathcal{N}_\mathcal{I}(.)\rvert$. Finally, performance comparisons w.r.t. the most related work are provided in Tab. 2 and some qualitative results in Fig. 2.

5. CONCLUSION

In this paper, we propose a deep hierarchical context network (DHCN) for image annotation. The method leverages two levels of contextual relationships; geometric and semantic. This is achieved by learning "end-to-end" the parameters of a deep context network whose architecture corresponds to the solution of an objective function that mixes a content criterion that maximizes the similarity between visually close contents, a context term which restores the similarity when content is versatile and a regularizer that smooths the similarity and helps providing a closed-form solution. Training the parameters of this deep context network, using a supervised SVM loss, makes it possible to learn the most influencing geometric and semantic contextual relationships for image annotation. Experiments conducted on the challenging ImageCLEF benchmark, show that the proposed DHCN substantially enhances the performances of image annotation compared to shallow context-free as well as deep context networks with handcrafted or learned (geometric only) contexts. As a future work, we are currently investigating other priors on geometric and semantic relationships in order to further enhance the performances of image annotation.

6. REFERENCES

[1] M. Jiu and H. Sahbi, "Semi supervised deep kernel design for image annotation," in *ICASSP*, 2015.

[2] V. N. Murthy, S. Maji, and R. Manmatha, "Automatic image annotation using deep learning representations," in *ICMR*, 2015, pp. 603–606.

[3] R. Wang, Y. Xie, J. Yang, L. Xue, M. Hu, and Q. Zhang, "Large scale automatic image annotation based on convolutional neural network," *JVCIR*, vol. 49, pp. 213–224, 2017.

[4] M. Jiu and H. Sahbi, "Nonlinear deep kernel learning for image annotation," *IEEE Transactions on Image Processing*, vol. 26(4), 2017.

[5] J. Zhang, Y. Mu, S. Feng, K. Li, Y. Yuan, and C.-H. Lee, "Image region annotation based on segmentation and semantic correlation analysis," *IET Image Processing*, vol. 12, no. 8, pp. 1331–1337, 2018.

[6] Q. Cheng, Q. Zhang, P. Fu, C. Tu, and S. Li, "A survey and analysis on automatic image annotation," *Pattern Recognition*, vol. 79, pp. 242–259, 2018.

[7] Y. Liu, K. Wen, Q. Gao, X. Gao, and F. Nie, "SVM based multi-label learning with missing labels for image annotation," *Pattern Recognition*, vol. 78, pp. 307–317, 2018.

[8] L. Zheng, Y. Yang, and Q. Tian, "SIFT Meets CNN: A Decade Survey of Instance Retrieval," *IEEE TPAMI*, vol. 40, no. 5, pp. 1224–1244, 2018.

[9] M. Jiu and H. Sahbi, "End-to-end deep kernel map design for image annotation," in *2020 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2020, pp. 1546–1550.

[10] P. K. Bhagat and P. Choudhary, "Image annotation: Then and now," *Image and Vision Computing*, vol. 80, pp. 1–23, 2018.

[11] M. Jiu and H. Sahbi, "Deep kernel map networks for image annotation," in *ICASSP*, 2016.

[12] H. Sahbi and X. Li, "Context-based support vector machines for interconnected image annotation," in *ACCV*, 2011, pp. 214–227.

[13] R. Salakhutdinov and G. Hinton, "Learning a nonlinear embedding by preserving class neighbourhood structure," *Journal of Machine Learning Research*, vol. 2, pp. 412–419, 2007.

[14] R. Hadsell, S. Chopra, and Y. LeCun, "Dimensionality reduction by learning an invariant mapping," in *CVPR*, 2006, pp. 1735–1742.

[15] M. Jiu, C. Wolf, G. Taylor, and A. Baskurt, "Human body part estimation from depth images via spatially-constrained deep learning," *Pattern Recognition Letteres*, vol. 50, pp. 122–129, 2014.

[16] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society. Series B*, vol. 58, no. 1, pp. 267–288, 1994.

[17] L. Jacob, G. Obozinski, and J. P. Vert, "Group lasso with overlap and graph lasso," *ICML*, 2009.

[18] F. Bach, R. Jenatton, J. Mairal, and G. Obozinski, "Structured sparsity through convex optimization," *Statistical Science*, vol. 27, no. 4, pp. 1–27, 2011.

[19] M. Belkin, P. Niyogi, and V. Sindhwani, "A geometric framework for learning from labeled and unlabeled examples," *JMLR*, vol. 7, pp. 2399–2434, 2006.

[20] M. Jiu and H. Sahbi, "Laplacian deep kernel learning for image annotation," in *ICASSP*, 2016.

[21] Y. LeCun, L. Botto, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[22] G. Hinton, S. Osindero, and Y-W. Teh, "A fast learning algorithm for deep belief nets," *Neural Computation*, vol. 18(7), pp. 1527–1554, 2006.

[23] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *NIPS*, 2012.

[24] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, pp. 436–444, 2015.

[25] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *2016 IEEE CVPR*, Jun 2016, pp. 770–778.

[26] M. Defferrard, X. Bresson, and P. Vandergheynst, "Convolutional Neural Networks on Graphs with Fast Localized Spectral Filtering," in *NIPS*, 2016.

[27] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," in *ICLR*, 2017.

[28] A. Ortega, P. Frossard, J. Kovacevic, J. M. Moura, and P. Vandergheynst, "Graph Signal Processing: Overview, Challenges, and Applications," *Proceedings of the IEEE*, vol. 106, no. 5, pp. 808–828, 2018.

[29] Z. Wu, S. Pan, F. Chen, Long G., C. Zhang, and P. S. Yu, "A comprehensive survey on graph neural networks," *CoRR*, vol. abs/1901.00596, 2019.

[30] C.-W. Lee, W. Fang, C.-K. Yeh, and Y.-C. Wang, "Multi-Label Zero-Shot Learning with Structured Knowledge Graphs," in *CVPR*, 2018, pp. 1576–1585.

[31] B. Knyazev, X. Lin, M. Amer, and G. Taylor, "Image classification with hierarchical multigraph networks," in *BMVC*, 07 2019.

[32] H. Sahbi, J.-Y. Audibert, and R. Keriven, "Context-dependent kernels for object classification," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, pp. 699–708, 2011.

[33] H. Sahbi, "Explicit context-aware kernel map learning for image annotation," in *ICVS*, 2013.

[34] M. Jiu, H. Sahbi, and L. Qi, "Deep Context Networks for Image Annotation," in *Proceedings - International Conference on Pattern Recognition*, 2018, pp. 2422–2427.

[35] M. Jiu and H. Sahbi, "Deep representation design from deep kernel networks," *Pattern Recognition*, vol. 88, pp. 447–457, 2019.

[36] C.-C. Chang and C.-J. Lin, "Libsvm: A library for support vector machines," *ACM Transactions on Intelligent Systems and Technology*, vol. 2, pp. 1–27, 2011.

[37] K. Chatfield, K. Simonyan, A. Vedaldi, and A. Zisserman, "Return of the devil in the details: Delving deep into convolutional nets," in *BMVC*, 2014.

[38] M. Varma and B. Babu, "More generality in efficient multiple kernel learning," in *ICML*, 2009.

[39] J. Zhuang, I. Tsang, and S. Hoi, "Two-layer multiple kernel learning," in *ICML*, 2011, pp. 909–917.