Deep Manifold Alignment for Mid-grain Sketch based Image Retrieval

Tu Bui[1], Leonardo Ribeiro[2], Moacir Ponti[2], and John Collomosse[1]

[1] University of Surrey — Guildford, Surrey GU2 7XH, UK
{t.bui,j.collomosse}@surrey.ac.uk
[2] Universidade de São Paulo — São Carlos/SP 13566-590, Brazil
{ leonardo.sampaio.ribeiro,ponti}@usp.br

Abstract. We present an algorithm for visually searching image collections using free-hand sketched queries. Prior sketch based image retrieval (SBIR) algorithms adopt either a category-level or fine-grain (instance-level) definition of cross-domain similarity — returning images that match the sketched object class (category-level SBIR), or a specific instance of that object (fine-grain SBIR). In this paper we take the middle-ground; proposing an SBIR algorithm that returns images sharing both the object category and key visual characteristics of the sketched query without assuming photo-approximate sketches from the user. We describe a deeply learned cross-domain embedding in which 'mid-grain' sketch-image similarity may be measured, reporting on the efficacy of unsupervised and semi-supervised manifold alignment techniques to encourage better intra-category (mid-grain) discrimination within that embedding. We propose a new mid-grain sketch-image dataset (MidGrain65c) and demonstrate not only mid-grain discrimination, but also improved category-level discrimination using our approach.

Keywords: SBIR · Manifold alignment · Visual Search.

1 Introduction

Free-hand sketch offers an intuitive and convenient query modality for visual search when a photographic sample of the desired content is unavailable. Yet, matching sketches and photographs is challenging; sketches are salient abstractions frequently drawn from canonical viewpoints, caricaturing objects, and introducing non-linear deformations [4, 21]. Recently deep neural networks, in particular multi-branch (triplet) networks, have proven effective in learning a mapping across these two domains for sketch based image retrieval (SBIR). Such approaches typically fall into either of two camps according to granularity at which matching is performed: 1) category (object-level) SBIR in which a sketched query of a given object (e. g. a cat) should return images containing that object (e. g. cats) [1–3]; 2) fine-grain (instance-level) search in which a detailed sketch of a specific object (e. g. a shoe) should return only that specific shoe [21, 16]. Whilst both bodies of work have made significant advances in cross-domain (sketch-photo) matching, arguably neither provides a model for practical SBIR. Category-level matching is analogous to sketched object classification, suggesting

that the need for a sketch could be obviated simply by substituting a coarse-grain label (e. g. text keyword) as query. Conversely, instance-level search requires unrealistic photo-approximate recall of fine-grain object detail within the sketched query — unreasonable due to both the limitations of human visual recall and typical depictive skill of users [4]. Furthermore it is challenging to obtain large quantities of fine-grain annotated training data.

(a) Instance-level matching in Sketchy [21]: (left most) sketches; (right most) relevant images; (middle) other images in the same categories.

(b) Our MidGrain65c dataset.

Fig. 1. Mid-grain matching of sketches to photographs; retrieved images match both object class and exhibit key visual characteristics of the sketch without demanding fine-grain, instance-level matching of a specific sketched object (per [21, 28]).

This paper proposes an intermediate level of granularity ('mid-grain') for cross-domain matching, in which the SBIR algorithm recalls images sharing both the object category and key visual characteristics of the sketched query (Fig. 1) but without requiring a precise sketch. For example, a sketch of an object in particular pose configuration returns similar objects in similar poses (e. g. front or side profiles of an elephant); or, a sketched sub-part of an object (e. g. piano keys) prioritizes recall of images dominated by that object part over images of the

whole object. Specifically we explore unsupervised and semi-supervised manifold alignment techniques to enhance the ability to perform mid-grain discrimination in a metric space, using a novel pooled sampling to select training examples for a triplet-loss network, using only category level annotation. Refinement of this embedding for mid-grain discrimination is performed iteratively, through intra-category clustering and correspondence (in some experimental configurations, using a small amount of fine-grained annotation) enabling sampling of hard positive/negatives from these clusters to drive refinement of the triplet network. We demonstrate that this process significantly improves not only the mid-grain discrimination, but also the category-level discrimination capability of the resulting embedding. To evaluate the ability of the trained network to perform mid-grain SBIR we collected a mid-grain annotated test set (*MidGrain65c*), and release this as a secondary contribution to our work.

2 Related Work

Early sketch based image retrieval (SBIR) algorithms tackled sketch-image matching as an optimization; fitting the sketch as a deformable model to image content and deriving rankings from the support evidenced for the sketched structure. Scalable approaches to SBIR began to emerge in the late-2000s, adapting gradient feature and dictionary learning approaches (popularised in photographic visual search) to SBIR. Notably, the Bag of Visual Words (BovW) paradigm was extended to SBIR using the Gradient-Field HoG (GF-HoG) [10, 11], Structure Tensor [7] and SHoG [6] descriptors all of which encode structure local to sparse key-points sampled from sketched strokes and Canny edge-pixels detected in images. Several BoVW indexing strategies were explored in [12, 1] including those fusing additional modalities such as colour, or semantic object labels. Mindfinder [23] used Chamfer matching to match sketched strokes to edge-lets extracted from images under an efficient indexing scheme. Indexing of mid-level sparse features were also explored through HELO [19] and key-shapes [20]. Whilst performance enhancements were achieved e. g. by substituting more perceptually inspired edge detectors for Canny [15] in the pre-processing, recent years have seen more significant advances through the use of deep convolutional neural networks (CNNs) [13] to learn the search embedding. CNNs were initially explored in the context of sketched object classification [29] through Sketch-A-Net; a truncated form of AlexNet [13]. Although such models can serve as feature extractors for SBIR, significant improvements in accuracy can be delivered through use of multi-branch (contrastive- or triplet-loss) networks. Such networks learn a cross-domain embedding by bringing together matching sketch-image (positive) pairs and pushing apart non-matching (negative) pairs within the learned embedding. Fully siamese triplet networks were explored for fine-grain SBIR in [28], and perform well for instance-level retrieval on a dataset of single object class (e. g. shoes or chairs). Heterogeneous triplet i. e. partial weight sharing networks in which some (or none) of the weights are shared across the sketch (anchor) and image (positive/negative) branches of the triplet network enables independent functions to be learned in order to map the disparate sketch and image domains

to a joint embedding for improved accuracy. Quadruplet networks [22] have also recently been explored, as have improved hard sampling strategies for triplet selection and assymetric feature matching [25]. Optimal weight sharing schemes and network architectures were studied extensively for category-level [2, 3] and fine-grain [21] on the Flickr15k [11] and Sketchy benchmarks respectively. Under the latter, the test set is formed by presenting image to human participants and inviting them to sketch the content. During retrieval, the only image considered correct is that originally used to derive the sketch (instance-level search) [21]. To the best of our knowledge, no prior work explores the training or evaluation of models given a mid-grain definition of similarity.

3 Deep Representation for Mid-grain Similarity

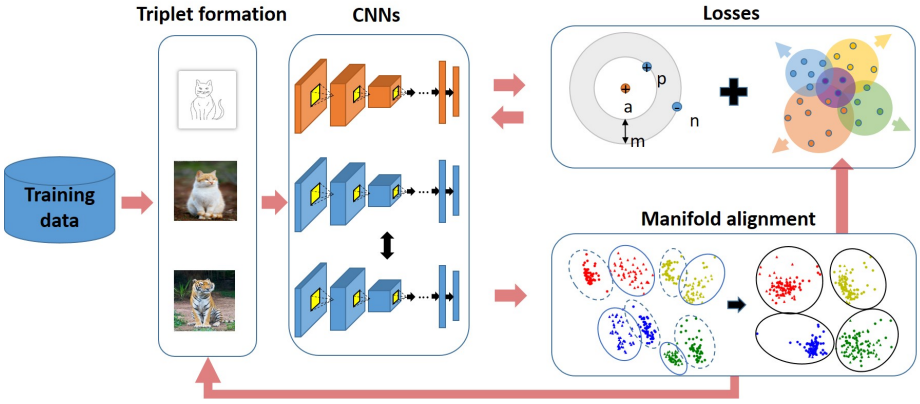


Fig. 2. An illustration of our mid-grain SBIR network.

We learn a cross-domain embedding for sketch-image matching using a triplet CNN (convnet) adopting a high performing network architecture from the set of variants proposed in [3]. The network chosen is a fully unshared (heterogeneous) triplet network with GoogleNet Inception-v1 backbone, shown to achieve state of the art category-level SBIR performance (53.26% mAP on Flickr15k[11]). Our core contribution is a novel method for selecting exemplar triplets i. e. anchor (query sketch), positive (+) and negative (-) images to form training tuples using a novel pooled sampling approach that yields significant improvement not only in category-level SBIR but uniquely enables also mid-grain discrimination in SBIR ranking (Fig. 2). We first provide an overview of the sampling process, then detail each step in the sub-sections 3.1-3.4.

Our training methodology is designed for the current scenario where very little instance-level data are available for training SBIR at fine-grain level. The positive and negative sketch-image pairs are therefore, be formed at class-level; however we propose to select only meaningful sketch-image pairs to feed the training network. First, two independently trained embeddings are initialized

for the sketch (anchor) and image (+/-) branches respectively, using pre-trained (GoogLeNet/ImageNet) weights refined by a few training epochs under classification (softmax) for the sketch and image data. These embeddings form are refined over subsequent training epochs. For each epoch, a set of triplets are sampled from a joint dataset of images and corresponding sketches grouped by object class (see Sec. 4.1 for dataset details). The training set is the union of several smaller sets, each sampled independently per object class as follows.

First, the two manifolds for the sketch and image representation of the object class are aligned using either an unsupervised or semi-supervised (with a small amount of fine-grain annotation) data. We evaluate the performance of three alignment approaches for this purpose. Next, unsupervised clustering is performed over the aligned distributions to characterize the intra-category (mid-grain) variation by pooling similar content. An anchor sketch and positive image are sampled from one resultant pool, used a stochastic sampling technique (data closer to a cluster centre is more likely to be selected). A similar stochastic sampling is applied to choose a negative image from a completely different category that has undergone similar pooling. By successive selection of triplets in this manner and training under a variant of the magnet loss function, the network weights are refined, which then form the embeddings for subsequent alignment, pooling and training iterations. We now explain each of these alignment, pooling and training processes in greater detail.

3.1 Sketch-Image manifold alignment

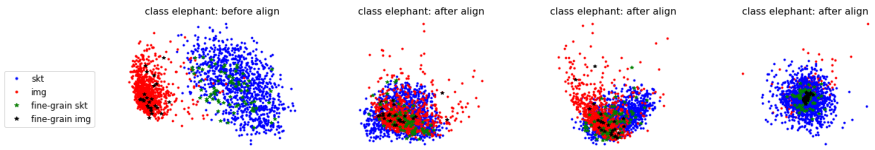
Consider a training set $\mathcal{D}_C = \{X_C, Y_C\}$ for a single object class C compromising N_S sketches $X_C = \{x_1, x_2, ..., x_{N_S}\}$ and N_I images $Y_C = \{y_1, y_2, ..., y_{N_I}\}$ (for simplicity we drop C from now on in all subsequent math notations unless otherwise stated). Supposed a subset of the training data has instance-level labels i. e. correspondence. For simplicity of exposition we assume the correspondence is one-to-one although this need not to be the case (see subsec. 3.1). Denote this subset $\mathcal{D}' = \{(x'_i, y'_i)\}_{i=1}^{M} \in \mathcal{D}$ where $M \ll \min(N_S, N_I)$. Denote $\mathcal{T} = \{f, g\}$ the parametrized embedding function that projects $\mathcal{D}$ into a P-dimensional embedding space – $U = f(X; \Theta_S) \in \mathbb{R}^{N_S \times P}$ and $V = g(Y; \Theta_I) \in \mathbb{R}^{N_I \times P}$ ($\mathcal{T}(.)$ is a triplet convnet in this work). We wish to align U with V in a common manifold for cross-domain similarity analysis to be implemented in the next step (subsec. 3.2). Note that finding a common embedding is also the ultimate objective of $\mathcal{T}(.)$ and our goal is to assist the search for the best embedding through selection of appropriate data (positive/negative pairs) to feed to the network under our subsequent pooled sampling step. We consider three approaches:

Unsupervised warping with PCA This naive approach does not require any sketch-image correspondence (i. e. fine-grain annotation). The approach assumes that sketches and images of a given category have the same distribution (mean, variance) in the latent (Mahalanobis) space. We employ PCA to derive Eigen decomposition for sketch and image representations, then warp the sketch representations U to have the same mean and variance as the image representations

V.

$$u_i := (u_i - \mu_S)E_S\Sigma_S^{-1/2}\Sigma_I^{1/2}E_I^T + \mu_I, \quad i = 1, 2, ..., N_S \tag{1}$$

where (μ_S, E_S, Σ_S) and (μ_I, E_I, Σ_I) are the mean, eigenvectors and eigenvalues of the sketches and images respectively. Fig. 3(b) shows the warping effects on an example category.



(a) before alignment (b) PCA (c) LTM (d) MA

Fig. 3. Alignment of two distributions for a single category. Sketch and image embeddings were captured after the first training iteration. Embedding dimension is 256-D originally, visualized in 2-D using PCA.

Learning a transformation matrix (LTM) We learn a linear transformation $\{W \in \mathbb{R}^{P\times P}, b \in \mathbb{R}^{1\times P}\}$ to warp $U' \rightarrow V'$ then apply it on the larger set $U \rightarrow V$. Since the fine-grain training set $\mathcal{D}'$ is small, regularization on W is needed to combat overfitting. Concretely we wish to optimize:

$$\arg\min_{W,b} \frac{1}{2M}\sum_{i=1}^{M} ||u_i'W + b - v_i'||^2 + \frac{\lambda}{2}||W - I||^2 \tag{2}$$

where λ is weight of the regularization term. W is forced closed to the identity matrix I since U and V are in a prospective joint embedding space. Note that we do not penalize the bias term b and tolerate free translation between the two distributions.

Eqn. 2 is solved using e. g. gradient descent. After learning (W, b), the sketch representations U is transformed to match with the images (Fig. 3(c)):

$$u_i := u_iW + b, \qquad i = 1, 2, ..., N_S \tag{3}$$

Manifold alignment (MA) Manifold alignment [26] assumes the two distributions have a similar underlining manifold. The approach aims to learn mapping functions $(F_S(.), F_I(.))$ to project the distributions (U, V) into a common space where not only correspondence but also local geometry are preserved. There are linear and non-linear approaches however we found the linear method more robust in our experiments. Additionally, the linear method produces explicit linear transformation matrices – $F(x) = x\mathcal{F}, \quad \mathcal{F} \in \mathbb{R}^{P\times d}$ where d is the dimension of the latent space) – therefore can be applied on unknown data. On the other

hands, there is no closed form for $(F_S(.), F_I(.))$ in the non-linear case. For the linear approach, the local geometry is preserved according to cost:

$$C_1(F_S, F_I) = \sum_{i,j}^{N_S} ||F_S(u_i) - F_S(u_j)||^2 W_{i,j}^S + \sum_{i,j}^{N_I} ||F_I(v_i) - F_I(v_j)||^2 W_{i,j}^I \quad (4)$$

where $W^S \in \mathbb{R}^{N_S \times N_S}$ and $W^I \in \mathbb{R}^{N_I \times N_I}$ are the pair-wise similarity matrices for sketch and image sets. An adjacency heat-map kernel defines W^S and W^I, e. g.:

$$W_{i,j}^S = \begin{cases} e^{-||u_i - u_j||^2} & \text{if } u_j \in \text{k nearest neighbour of } u_i \\ 0 & \text{otherwise} \end{cases} \quad (5)$$

The inter-domain correspondence is also preserved according to cost function:

$$C_2(F_S, F_I) = \sum_{i \in [1,N_S], j \in [1,N_I]} ||F_S(u_i) - F_I(v_j)||^2 W_{i,j}^{S,I} \quad (6)$$

where $W^{S,I} \in \mathbb{R}^{N_S \times N_I}$ is the inter-adjacency similarity matrix between the sketch and image sets. If a sketch-image pair is a known correspondence their similarity score is set to high, and low if their correspondence is unknown:

$$W_{i,j}^{S,I} = \begin{cases} 1 & \text{if } (u_i, v_j) \in \mathcal{D}' \\ 0 & \text{otherwise} \end{cases} \quad (7)$$

Note from the way eqn. 6 is formulated, manifold alignment does not explicitly require one-to-one correspondence. The final cost function integrates both intra- and inter-loss:

$$C(F^{(S)}, F^{(I)}) = C_1(.) + \alpha C_2(.) \quad (8)$$

We set $\alpha = 2.0$ to stress the importance of the fine-grain set in the total loss (the value of α is not so sensitive to performance as demonstrated empirically in Fig. 4c). We refer to [26] for the way to solve eqn. 8. Fig. 3(d) visualizes a warping example for a representative category using this method.

3.2 Intra-Category Clustering

Following the alignment of sketch and image domains for each category, the combined data is clustered into blobs of similar sketches and images. We experimented with four unsupervised clustering techniques considering their abilities to automatically select the number of clusters for each category.

***k*-means** divides the sample set into K disjoint clusters, repeatedly update the clusters minimizing sum of square distance between samples and its centroids. We initialized the centroids using "kmeans++" and set number of clusters fixed at K=5 (by inspection, based on typical number of human-separable appearance variants within each category).

Gaussian Mixtures (GMM) We followed the GMM fitting protocol of [18] to determine the number of clusters automatically by penalizing number of free

parameters (thus number of clusters) in the mixture. This typically results in 3-4 clusters for each category.

Mean Shift [5] widely used in clustering, segmentation and tracking applications. It is a recursive non-parametric technique for locating maxima of a density function that approximates the sample set.

DBSCAN [8] locates data points with the highest neighbourhood density then expands clusters from them. The algorithm does not require prior knowledge of cluster number.

3.3 Magnet loss

Softmax loss has been shown effective at sketch classification [29] and in [21, 22, 3] as a step in training SBIR embeddings. Yet we observe for mid-grain SBIR that softmax loss causes the intra-category sketch-time distribution to narrow, frustrating pooling. We adopt an approach similar to magnet clustering [17] but adapted to retrieval across domains rather than a single-domain classification. Magnet loss maintains a set of clusters within each category and minimizes the accumulated distance from the data points to their own centroids, as opposed to one single point in softmax loss.

$$\mathcal{L}(\Theta) = \frac{1}{N}\sum_{n=1}^{N}\{-\log\frac{e^{-\frac{1}{2\sigma^2}||z_n-\mu(z_n)||^2-\beta}}{\sum_{c\neq C(z_n)}\sum_{k=1}^{K}e^{-\frac{1}{2\sigma^2}\left|\left|z_n-\mu_k^c\right|\right|^2}}\}_+ \tag{9}$$

where μ_z is centroid of the cluster containing z, μ_k^c is centroid of cluster k in class c, $C(z)$ is category of z, σ^2 is variance of all sample z away from their respective centroid μ_z, $\beta \in \mathbb{R}$ is the threshold for acceptable distance between z and its centroid. The original magnet loss [17] is adapted for cross-domain retrieval as follows:

1. True cluster centroids μ_z and μ_k^c are used instead of approximating them within a mini-batch. Due to memory constraints the number of samples per cluster to be fed into each mini-batch is quite small. Therefore, approximating the centroids using just the samples in a mini-batch might lead to inaccurate results. Instead we pre-compute the centroids each time the clusters are updated, then feed them back to the training as fixed vectors (see subsec. 3.4).
2. There is no constraint on the number of clusters per category. We used various clustering techniques (subsec. 3.2), many of them with auto-selection of the cluster number, instead of just $k-$means with fixed k clusters [17].
3. If the clustering process is implemented on the data-aligned space, the cluster means (centroids) must be subsequently warped back to the embedding space. It is possible since the linear transformation of manifold alignment is reversible. In cases of warping using PCA or LTM, the sketch space is warped while the image space is kept fixed. In order to unify the cross-domain learning objectives we do not revert the centroids back to the sketch space.

3.4 Triplet Formation

Following data alignment (subsec. 3.1) and clustering (subsec. 3.1), we obtain a set of clusters (intra-category pools) along with their centroids for each class. The sketches and images of the same pool are candidates for positive pairs; whilst the ones in the nearest impostor pool are negative candidates. The list of centroids is used as the learning targets for magnet loss (eqn. 9) in the next training iteration.

Triplet is formed as follows:

1. Sample a seed class C as a uniform distribution.
2. Sample a cluster $l \sim p_c(l)$.
3. Sample a sketch $x^s \sim p_c^l(x^s)$ and a positive image $x_+^I \sim p_c^l(x_+^I)$.
4. Sample a negative image in the nearest impostor cluster l' – $x_-^I \sim p_c^{l'}(x_-^I)$.

where $p_c(l)$ is size of cluster l in class C. $p_c^l(x)$ is a function of probability that sample x belongs to cluster l. For example, if the clustering method is GMM, $p_c^l(x)$ is the probability density function of x given cluster l. In other cases, we used the distance heat map to represent $p_c^l(.)$.

Data augmentation – We apply the following augmentation methods to enrich population of the training images and sketches: random crop (from 256×256 to 224×224), random rotation within a small range of [-5,5] degrees and scaling with random ratio in range [0.9,1.1]. Random flip is not applied to preserve object viewpoint. Uniquely for sketches, we randomly discard up to 10% of stroke number. Data augmentation and sketch rendering are implemented on-the-fly in parallel with the main learning stream for speed efficiency.

4 Experiments

We first describe the training and test datasets in subsec. 4.1, and evaluate algorithm configurations to determine our optimal models in subsec. 4.2. Finally subsec. 4.3 shows performance of our proposed model in comparison with other approaches.

4.1 Datasets

	QuickDraw [9]	ImageNet [13]	Sketchy [21]	Common65c-coarse	Common65c-fine
Class number	345	1000	125	65	65
# sketches	50M	0	75K	65K	4.7K
# images	0	1.2M	12.5K	65K	1.6K

Table 1. The Common65c dataset is formed using sketches and images from ImageNet, QuickDraw and Sketchy.

As the *training* of our proposed approach involves clustering samples within the same categories, it is necessary to have large number of class-level sketches

and images, plus (for LTM/MA) a small set of fine-grain data. QuickDraw [9] and ImageNet [13] are currently the largest datasets of sketches and images respectively, while Sketchy [21] is the largest instance-level SBIR dataset. We therefore intersect the category lists of these three datasets and obtain 65 common object categories from which we form a training set called Common65c. Specifically, Common65c consists of two subsets: a class-level subset, *Common65c-coarse*, and a instance-level subset, *Common65c-fine*. Common65c-coarse has 65k sketches from Quickdraw and 65k photo images from ImageNet, while Common65c-fine has 4680 sketches and 1560 images from Sketchy (Tab. 1). Note that we purposely restrict the size of the Common65c-fine set to just 24 images and 72 sketches per category (one image has 3 sketch correspondents) so that it is negligible against the Common65c-coarse. That would fit our original goal of building a semi-supervised mid-grain SBIR model.

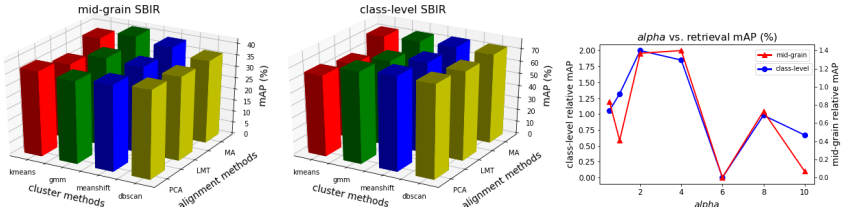
A mid-grain *evaluation* dataset is also needed. To obtain the sketch set we sampled from QuickDraw 200 random sketches for each of the 65 categories, holding out data already used in the Common65c-coarse training set. A set of 138 sketches were sampled to form a balanced evaluation set manually selecting distinct views or sub-types of objects within each object class. Each sketch encodes a single mid-grain variant of an object.

To obtain the set of images corresponding to the sketch queries we scraped images through text keyword search for the 65 category names of Common65c on Adobe Stock image search. We chose this repository over Flickr, Google and Bing to avoid overlap with the ImageNet training set. Crowd annotation was used to select 1247 strong matches from the 500 images per category downloaded. We also added random 'distractor' images from Adobe Stock to form a 100k image corpus. This new dataset is called *MidGrain65c*. Several examples of sketch and corresponding images are shown Fig. 1.

We also created smaller datasets by sub-sampling 12 categories out of 65s, namely Common12c and MidGrain12c. These datasets were used to evaluate the clustering and alignment methods in the next section. All datasets are released as a further contribution.

4.2 Clustering and alignment methods

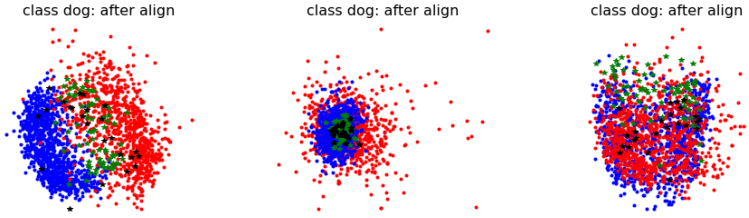
We experimented with the data alignment (PCA, LTM and MA) and clustering methods ($k-$means, GMM, Mean Shift, DBSCAN) of subsec. 3.1-3.2. Training and evaluation were implemented on the Common12c and MidGrain12c sets respectively. Performance of these techniques is shown in Fig. 4(a-b). At mid-grain level (Fig. 4(a)), MA-GMM has the highest performance at 41.1% mAP, while LTM-KMeans performs the worst at 33.2%. At class-level (Fig. 4(b)), LTM-KMeans again under-performs the others at 61.9% but the highest accuracy is achieved with PCA-MeanShift at 76.3%. Other methods that rely on PCA alignment also obtain good results e. g. 74.2% for PCA-GMM, 75.9% for PCA-DBSCAN, thanks to its generic unbiased (and unsupervised) mechanism. It is opposite to the mid-grain case where MA dominantly outperforms others. Interestingly, the supervised LTM methods perform no better than PCA. Fig. 5



(a) mid-grain. (b) class-level. (c) Tuning α in MA.

Fig. 4. Comparing SBIR performance (mAP over MidGrain12c) for clustering and data alignment methods for (a) mid-grain and (b) category-level retrieval. (c) Effects of α in eqn. 8 on MA performance.

shows a failed example of LTM where the fine-grain set is aligned but the majority of sketches and images are still separated. This is usually not the case for MA since it not only aligns correspondents but also respect local geometry.



(a) LTM (b) MA (c) PCA

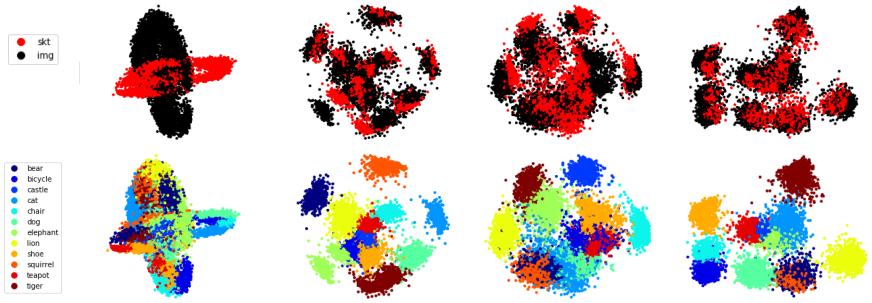
Fig. 5. A failure case of (a) LTM, as compared with (b) MA and (c) PCA.

Additionally, Fig. 4(a-b) indicates no correlation between mid-grain mAP and class-level mAP i. e. being the most superior at mid-grain level does not guarantee the same for category-level and vice versa. It is probably due to the trade off between discrimination (favoured in fine-grain SBIR) and generalization (preferred in category-level SBIR). This further encourages studies of mid-grain SBIR which comfortably sits between the two.

4.3 Baseline comparison

We selected the best performing model in the previous experiments (**MA-GMM**) and trained it on the full dataset (Common65c). The following baselines were compared against:

SS-triplet-HM, standard triplet network with single staged training and hard-mining. The same network architecture as MA-GMM is employed (no-share 256-D InceptionV1) and the weight is initialized using the pretrained ImageNet model [24]. We implemented online hard-negative mining where the closest negative image within a mini-batch is selected for each anchor sketch. The whole Common65c dataset is used in training although the fine-grain labels of the subset Common65c-fine are not being used.



(a) Before training (b) MA-GMM (c) SS-triplet-HM (d) SS-contrast-HM

Fig. 6. PCA distribution of the Common12c data before and after training by domains (top row) and by categories (bottom row).

SS-contrast-HM, standard contrastive-loss network with single staged training and hard-mining. Otherwise identical to SS-triplet-HM.

MS-reg-HM, multi-staged regression network with hard-mining proposed in Bui *et al.* [3]. The same architecture as MA-GMM is employed except the sketch and image branches are partially shared from block inception-4e.

Sketchy [21], fine-grain triplet-based network trained on the 75K Sketchy dataset. We used this publicly available model as a standard-alone baseline. Note we did not fine-tune it on Common65c since (i) it is a coarse-grain dataset and (ii) its 65 categories are a subset of the larger 125 Sketchy categories.

All other settings, unless specified otherwise, are kept the same.

Tab. 2 compares performance of these approaches on MidGrain65c at mid-grain and class-level. MS-REG-HM performs better than SS-triplet-HM which in turn is superior to SS-contrast-HM. The Sketchy model surprisingly has the lowest performance even though it was trained on a much more diverse and fine-grained dataset (we note that the sketch queries in MidGrain65c are originally from QuickDraw which is less clean than Sketchy due to the way QuickDraw was created). Sketchy also suffers a severe drop in performance when noisy distracting images are added to the benchmark, which shows its lack of generalization to "images in the wild". On top of that, our proposed approach MA-GMM outperforms the second-best by 6% and proves to be more robust in presence of distracting images. Note that MA-GMM needs just one training step as opposed to three stages in MS-reg-HM. It also employs a sub-optimal sharing configuration (no-share network) and does not directly use the fine-grain set to train its parameters. Fig. 7 shows representative SBIR results.

Fig. 6 visualizes the distributions of the training data Common12c before and after the networks were trained. Fig. 6(a) was created after the third epoch in which sketches and images were being pre-trained separately using softmax loss (which serves as weight initialisation for MA-GMM). Intra-category pooling is visible however the two domains were inter-mixed due to a lack of cross-domain training. Contrastive loss makes the distribution for each category more com-

(a) MA-GMM

(b) MS-REG-HM

(c) Sketchy

Fig. 7. Mid-grain SBIR results of several representative queries (class "church", "cat" and "bench"), including failure cases. Red and yellow bounding boxes depict non-relevant images; the later indicates images of the same classes as the queries.

Methods	mid-grain mAP (%)		class-level mAP (%)	
	w. distract	w/o distract	w. distract	w/o distract
MA-GMM	**42.10**	**48.40**	**65.31**	**79.17**
MS-reg-HM [3]	36.08	45.58	53.52	74.01
SS-triplet-HM	32.13	43.39	45.85	69.35
SS-contrast-HM	22.34	42.65	31.64	66.82
Sketchy[21]	12.86	39.72	12.65	47.82

Table 2. Mid-grain and class-level SBIRs of MA-GMM in comparison with other approaches, tested on MidGrain-65c with and without distracting images.

pact, reducing intra-category discrimination (Fig. 6(d)). The more flexible triplet loss makes for wider distributions (Fig. 6(c)), however several distributions were mixed up probably due to strict hard-negative mining being less effective against noisy data. MA-GMM brings more balance to the distributions, maintains the inter-category separation at the same time avoids squeezing the intra-category distance (Fig. 6(b)).

5 Conclusions

We report the first mid-grain SBIR algorithm; an unexplored topic fitting between object category and instance retrieval. We proposed a semi-supervised approach that utilizes mainly class-level datasets and a small quantity of fine-grain annotation combined with unsupervised intra-category clustering. We build upon the past success of triplet convnets for SBIR [21, 3] and the trend in visual search more broadly that targeted selection of triplets (e. g. hard-negative mining over conventional random sampling [25]) yields performance improvements. We go further, proposing a guided sampling scheme in which sketch-image representations within intra-category are aligned and pooled. We studied various data alignment and clustering strategies to determine the best combination (MA/GMM) for pooling. The whole process is integrated into a single staged end-to-end learning framework. We demonstrated our approach superior to other traditional methods on a newly created mid-grain dataset, MidGrain65c, by a 6% margin. Training time reduction is the main topic for future work. As manifold alignment and cluster updates are implemented on a regular basis, training needs to be frozen at the same frequency. Additionally, the requirement of a small amount of fine-grain training annotation (for best performance) is another limitation and an unsupervised approach that outperforms PCA is desirable. The need for this annotation narrows the diversity of our training set to 65 available categories. Another direction is developing attentive-models that focus on auto-detected regions of interest rather than the whole images. Recent work in this direction in broader image retrieval [27, 14] could be adapted.

Acknowledgments

This work was supported in part via an EPSRC doctoral training studentship (EP/M508160/1) and in part by UGPN/RCF 2017, FAPESP (grants 2016/16111-4, 2017/10068-2 and 2013/07375-0) and CNPq Fellowship (#307973/2017-4).

References

1. Bui, T., Collomosse, J.: Scalable sketch-based image retrieval using color gradient features. In: Proceedings of the IEEE International Conference on Computer Vision Workshops. pp. 1–8 (2015)
2. Bui, T., Ribeiro, L., Ponti, M., Collomosse, J.: Compact descriptors for sketch-based image retrieval using a triplet loss convolutional neural network. Computer Vision and Image Understanding **164**, 27–37 (2017). https://doi.org/http://dx.doi.org/10.1016/j.cviu.2017.06.007, http://www.sciencedirect.com/science/article/pii/S1077314217301194
3. Bui, T., Ribeiro, L., Ponti, M., Collomosse, J.: Sketching out the details: Sketch-based image retrieval using convolutional neural networks with multi-stage regression. Computers & Graphics **71**, 77–87 (2018)
4. Collomosse, J.P., McNeill, G., Watts, L.: Free-hand sketch grouping for video retrieval. In: International Conference on Pattern Recognition (ICPR) (2008)
5. Comaniciu, D., Meer, P.: Mean shift: A robust approach toward feature space analysis. IEEE Transactions on pattern analysis and machine intelligence **24**(5), 603–619 (2002)
6. Eitz, M., Hays, J., Alexa, M.: How do humans sketch objects? ACM Trans. on Graphics (Proc. SIGGRAPH **31**(4), 44:1–44:10 (2012)
7. Eitz, M., Hildebrand, K., Boubekeur, T., Alexa, M.: A descriptor for large scale image retrieval based on sketched feature lines. In: Proc. SBIM. pp. 29–36 (2009)
8. Ester, M., Kriegel, H.P., Sander, J., Xu, X., et al.: A density-based algorithm for discovering clusters in large spatial databases with noise. In: Kdd. vol. 96, pp. 226–231 (1996)
9. Ha, D., Eck, D.: A neural representation of sketch drawings. arXiv preprint arXiv:1704.03477 (2017)
10. Hu, R., Barnard, M., Collomosse, J.P.: Gradient field descriptor for sketch based retrieval and localization. In: Image Processing (ICIP), 2010 IEEE International Conference on. vol. 10, pp. 1025–1028 (2010)
11. Hu, R., Collomosse, J.: A performance evaluation of gradient field HOG descriptor for sketch based image retrieval. Computer Vision and Image Understanding **117**(7), 790–806 (2013). https://doi.org/10.1016/j.cviu.2013.02.005
12. Hu, R., James, S., Wang, T., Collomosse, J.: Markov random fields for sketch based video retrieval. In: Proceedings of the 3rd ACM conference on International conference on multimedia retrieval. pp. 279–286. ACM (2013)
13. Krizhevsky, A., Sutskever, I., Hinton, G.: Imagenet classification with deep convolutional neural networks. In: Advances in neural information processing systems (2012)
14. Laskar, Z., Kannala, J.: Context aware query image representation for particular object retrieval. In: Scandinavian Conference on Image Analysis. pp. 88–99. Springer (2017)
15. Qi, Y., Song, Y.Z., Xiang, T., Zhang, H., Hospedales, T., Li, Y., Guo, J.: Making better use of edges via perceptual grouping. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2015)
16. Qi, Y., Song, Y.Z., Zhang, H., Liu, J.: Sketch-based image retrieval via siamese convolutional neural network. In: Image Processing (ICIP), 2016 IEEE International Conference on. pp. 2460–2464. IEEE (2016)
17. Rippel, O., Paluri, M., Dollar, P., Bourdev, L.: Metric learning with adaptive density discrimination. arXiv preprint arXiv:1511.05939 (2015)

18. Roberts, S.J., Husmeier, D., Rezek, I., Penny, W.: Bayesian approaches to gaussian mixture modeling. IEEE Transactions on Pattern Analysis and Machine Intelligence **20**(11), 1133–1142 (1998)
19. Saavedra, J.M.: Rst-shelo: sketch-based image retrieval using sketch tokens and square root normalization. Multimedia Tools and Applications **76**(1), 931–951 (2017)
20. Saavedra, J.M., Barrios, J.M.: Sketch based image retrieval using learned keyshapes. In: Proceedings of the British Machine Vision Conference (2015)
21. Sangkloy, P., Burnell, N., Ham, C., Hays, J.: The sketchy database: learning to retrieve badly drawn bunnies. ACM Transactions on Graphics (TOG) **35**(4), 119 (2016)
22. Seddati, O., Dupont, S., Mahmoudi, S.: Quadruplet networks for sketch-based image retrieval. In: Proceedings of the 2017 ACM on International Conference on Multimedia Retrieval. pp. 184–191. ACM (2017)
23. Sun, X., Wang, C., Xu, C., Zhang, L.: Indexing billions of images for sketch-based retrieval. In: Proceedings of the 21st ACM international conference on Multimedia. pp. 233–242. ACM (2013)
24. Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A.: Going deeper with convolutions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1–9 (2015)
25. Tolias, G., Chum, O.: Asymmetric feature maps with application to sketch based retrieval. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. vol. 1, p. 4 (2017)
26. Wang, C., Mahadevan, S.: A general framework for manifold alignment. In: AAAI fall symposium: manifold learning and its applications. pp. 53–58 (2009)
27. Wei, X.S., Luo, J.H., Wu, J., Zhou, Z.H.: Selective convolutional descriptor aggregation for fine-grained image retrieval. IEEE Transactions on Image Processing **26**(6), 2868–2881 (2017)
28. Yu, Q., Liu, F., Song, Y.Z., Xiang, T., Hospedales, T.M., Loy, C.C.: Sketch me that shoe. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. IEEE (2016)
29. Yu, Q., Yang, Y., Song, Y.Z., Xiang, T., Hospedales, T.M.: Sketch-a-net that beats humans. In: Proceedings of the British Machine Vision Conference. IEEE (2015)