Academic Coupled Dictionary Learning for Sketch-based Image Retrieval

Dan Xu[♯], Xavier Alameda-Pineda[♯], Jingkuan Song[♯], Elisa Ricci[‡], Nicu Sebe[♯]
[♯]Department of Information Engineering and Computer Science, University of Trento, Trento, Italy
[‡]Fondazione Bruno Kessler (FBK), Trento, Italy and University of Perugia, Perugia, Italy
{dan.xu, xavier.alamedapineda, jingkuan.song, niculae.sebe}@unitn.it,elisa.ricci@fbk.eu

ABSTRACT

In the last few years, the *query-by-visual-example* paradigm gained popularity, specially for content based retrieval systems. As sketches represent a natural way of expressing a synthetic query, recent research efforts focused on developing algorithmic solutions to address the sketch-based image retrieval (SBIR) problem. Within this context, we propose a novel approach for SBIR that, unlike previous methods, is able to exploit the visual complexity inherently present in sketches and images. We introduce academic learning, a paradigm in which the sample learning order is constructed both from the data, as in self-paced learning, and from partial curricula. We propose an instantiation of this paradigm within the framework of coupled dictionary learning to address the SBIR task. We also present an efficient algorithm to learn the dictionaries and the codes, and to pace the learning combining the reconstruction error, the prior knowledge suggested by the partial curricula and the cross-domain code coherence. In order to evaluate the proposed approach, we report an extensive experimental validation showing that the proposed method outperforms the state-of-the-art in coupled dictionary learning and in SBIR on three different publicly available datasets.

CCS Concepts

•Information systems → Retrieval models and ranking; Image search; •Computing methodologies → Machine learning;

Keywords

Sketch-based image retrieval; self-paced and curriculum learning; dictionary learning

1. INTRODUCTION

Favored by the widespread diffusion of consumer touchscreen devices, recently sketch-based image retrieval became popular as an effective means of querying large image databases. Most prior works on SBIR [12, 8, 27, 28, 9, 29] focused on designing robust feature representations, common for both sketches and images, to allow a direct matching. Typically such approaches involve three steps: (i) extract edges from images so to approximate the sketch, (ii) compute a bag-of-words representation [9] of the sketches and of the resulting edge maps and (iii) finally query by matching the sketch to the edge maps in the common feature space. The major issue with this strategy is that the features are not learned to satisfy the final objective, *i.e.*, robust matching, but hand-crafted in advance under the hypothesis that the statistical distributions of the extracted image edges and the sketch edges are the same. In addition, such strategy must use the same features for both image and sketch modalities, while intuitively different features could better represent different modalities. It is thus unsurprising that more recent studies proposed learning methods able to deal with different feature representations, thus partially addressing the limitations of the classical strategy. A series of approaches based on dictionary learning (DL) [19] was proven especially useful for learning coupled sparse representations [42, 36, 13] from data coming from different modalities.

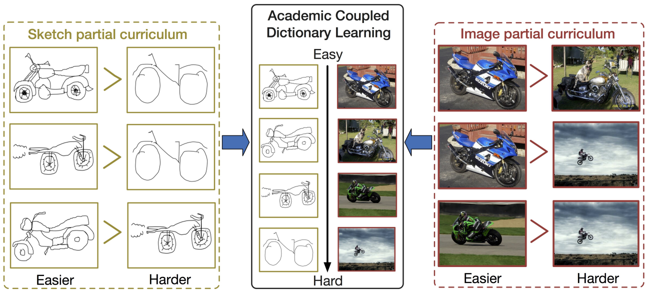
Sketch partial curriculum
Easier
Harder
Academic Coupled Dictionary Learning
Easy
Hard
Image partial curriculum
Easier
Harder

Figure 1: In real SBIR scenarios, humans can easily assess which of the images or sketches in a pair is easier to learn from, while providing a full order of the image/sketch set is chimeric. Academic coupled dictionary learning combines the flexibility of partial modality-specific curricula with the power of self-paced learning strategies to automatically construct a full sample learning order that evolves over time until all training samples are used to build a dictionary-based representation.

In parallel, curriculum learning (CL) [3] and self-paced learning (SPL) [18] strategies, have proven to be advantageous for various tasks in computer vision and multimedia, such as object tracking [32], event detection [16] or multimedia search [15]. Both CL and SPL work by constructing a sample learning order that depends on the inherent data complexity, such as to increase the chances of the learner to avoid local minima. Intuitively, SBIR would also benefit from such strategies, since the visual complexity across samples varies significantly. Indeed, natural images are characterized by cluttered backgrounds and objects-of-interest captured at different scales or various poses. Similarly, the visual complexity of sketches drawn by expert/non-expert shows notorious variations. Both CL

MM '16 Amsterdam, The Netherlands
© 2016 ACM. ISBN 123-4567-24-567/08/06. . . $15.00
DOI: 10.475/123_4

and SPL aim to provide a sample learning order in which the easiest samples are presented first. However, while in CL the order is provided by an expert, in SPL the algorithm extracts the order from the data points typically based on the representation power of the model being learned. Recently, Jiang *et al.* [17] demonstrated the advantages of combining CL and SPL within the framework of dictionary learning. The authors model the given curriculum as a sort of a prior for the pacing variable. However, there are two main limitations in this approach. Firstly, data are assumed to come only from a single domain, and hence the method is not specifically designed for multimedia tasks such as SBIR. Secondly and more important, a full curriculum (*i.e.*, complete order) must be specified. This limits the applicability of the method to small-medium scale problems, since the curriculum is usually designed by humans. Moreover, in many tasks (*e.g.*, SBIR), assessing the easiness of each sample (*e.g.*, image or sketch) may not be straightforward, and thus full curricula are naturally inappropriate. Finally, full curriculum strategies are not appropriate either for continuously growing datasets, since each sample must be included into the full order and this process can be tedious and highly time-consuming.

To tackle the aforementioned issues, in this paper we present a novel learning paradigm and propose an instantiation within the framework of dictionary learning so to address the SBIR task. In academia, the students learn from different sources (textbooks, courses, on-line, etc) and, more importantly, no single full curriculum is available, but many different partial curricula (one per source). Indeed, the learning order suggested by the professor will be different from the order in which a textbook presents the concepts. In our model, instructors provide prior knowledge in the form of a partial order of samples, while leaving students the freedom to decide the full curriculum according to their learning pace. We formalize this intuition and introduce academic learning, where different partial curricula influence the learning pace (see Figure 1). More precisely, we address the SBIR task with an instantiation of academic learning within the framework of coupled dictionary learning, leading to *academic coupled dictionary learning* (ACDL). Intuitively, our method learns a pair of image- and sketch-specific dictionaries, together with the sparse codes, enforcing the similarity between the codes of corresponding sketches and images. At each iteration, the representation power of the learned dictionaries, the code correspondence and the partial modality-specific curricula determine which samples to learn from. We extensively evaluate ACDL on three publicly available datasets, demonstrating improved retrieval performance with respect to previous cross-domain dictionary learning methods and state-of-the art SBIR approaches.

To summarize, the main contributions of this paper are: (i) We introduce the academic learning paradigm, so to effectively merge the self-pacing philosophy with modality-specific partial curricula and propose an instantiation within the framework of coupled dictionary learning, leading to ACDL so as to address the SBIR task; (ii) We derive an efficient algorithm to learn the modality-specific dictionaries and codes, assessing the learning pace from the dictionary representation power, the code correspondence and the a priori partial order from the curricula; (iii) We report an extensive experimental evaluation proving that ACDL outperforms previous coupled dictionary learning methods and other state-of-the-art SBIR approaches in three different publicly available datasets.

2. RELATED WORK

In this section we describe related work on (i) the addressed task, SBIR, (ii) the ground learning framework, cross-domain DL and (iii) the closest learning paradigms to academic learning: SPL and curriculum learning.

2.1 Sketch-based image retrieval

Early works on SBIR attempted to use generic descriptors, *i.e.*, features not designed for this application. Both global (*e.g.*, distribution of edge pixels [6], elastic contours [4]) and local (*e.g.*, SIFT [9]) feature representations were investigated. More recently, SBIR methods focused on designing *ad hoc* descriptors, *i.e.*, features able to properly describe sketches, which are sparse images containing strokes. For instance, Hu *et al.* [12] presented gradient field images and used them in combination with a bag-of-words approach for retrieving images from sketches. They also performed an extensive evaluation of these descriptors on a large dataset, named the Flickr15k dataset. Saavendra *et al.* [28] proposed a modified version of the HOG descriptor, called histogram of edge local orientations, to explicitly tackle the problem of sparsity arising when traditional HOG descriptors are applied to sketches. In [27] mid-level patterns called learned keyshapes were introduced for representing sketches. Sun *et al.* [31] described a large scale SBIR system based on features derived from edge pixels and Chamfer matching. Shape words features for SBIR were introduced in [38]. Lin *et al.* [22] proposed the 3D sub-query expansion approach: to improve the retrieval performance, 2D sketches are first converted into a 3D sketch model, which is then used to generate multi-view sketches acting as expanded sub-queries. Qi *et al.* [26] introduced a perceptual grouping framework to organize image edges into a meaningful structure and used it for generating human-like sketches useful for SBIR. Fine-grained SBIR was tackled in [21], where the emphasis is on modeling fine-grained characteristics of sketches to retrieve specific images within object categories, as opposite to traditional SBIR focusing on retrieving images of the same class. Cao *et al.* [5] introduced MindFinder, an interactive sketch-based image search engine where multimodal queries (sketch+text) are used for retrieval. A sketch-based system for manga image retrieval was described in [24]. However, none of these works proposed a cross-domain DL approach for SBIR.

2.2 Cross-domain dictionary learning

Dictionary learning approaches [19] have become popular in computer vision and multimedia, as they were shown to be effective in many tasks, such as image denoising [23] and video event detection [41]. When data from multiple domains are available, traditional DL approaches have been extended to benefit from the wealth of multimodal information. Yang *et al.* [42] proposed to learn a set of source-specific dictionaries from samples corresponding to different domains in a coupled fashion and applied it to image super-resolution. In [36] a semi-coupled DL scheme was introduced, where source-specific dictionaries were learned together with a mapping function which describes the intrinsic relationship between domains. Similarly, Huang and Wang [13] proposed a framework to simultaneously learn a pair of domain-specific dictionaries and the associated representations. Coupled DL approaches have been tested on the SBIR task both in [36] and [13]. Similarly, DL has been employed for other related tasks, *i.e.*, sketch-based 3D object retrieval [43] or sketch recognition [11]. However, none of these works considers the sample easiness while learning, *i.e.*, apply a SPL or a curriculum learning strategy, in order to construct robust representations.

2.3 Self-paced and curriculum learning

Self-paced [18] and curriculum [3] learning originate from the idea that models must be learned in an incremental fashion, using the easy samples before the difficult ones. Due to their generality, these techniques have been considered in a broad spectrum of learning algorithms, including matrix factorization [45, 17], clus-

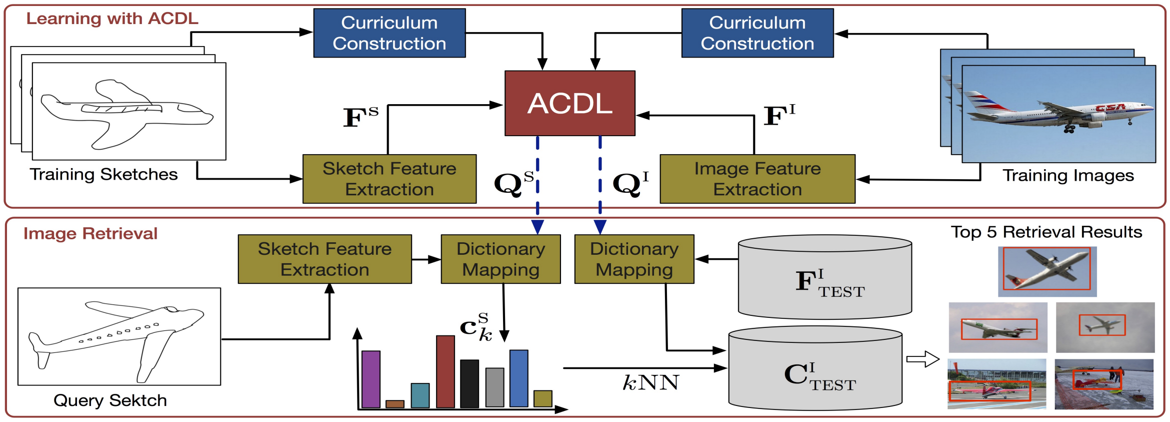
Learning with ACDL
Curriculum Construction
Curriculum Construction
ACDL
$\mathbf{F}^{\mathrm{S}}$
$\mathbf{F}^{\mathrm{I}}$
Training Sketches
Sketch Feature Extraction
$\mathbf{Q}^{\mathrm{S}}$
$\mathbf{Q}^{\mathrm{I}}$
Image Feature Extraction
Training Images
Image Retrieval
Sketch Feature Extraction
Dictionary Mapping
Dictionary Mapping
$\mathbf{F}^{\mathrm{I}}_{\mathrm{TEST}}$
Top 5 Retrieval Results
$\mathbf{c}^{\mathrm{S}}_k$
kNN
$\mathbf{C}^{\mathrm{I}}_{\mathrm{TEST}}$
Query Sektch

Figure 2: Overview of the proposed ACDL framework for sketch-based image retrieval.

tering [39], multi-task [25] and dictionary learning [34]. In the area of multimedia they have proved successful in many applications such as media retrieval [15] and event detection [16].

Among previous works, the two studies essentially related to this paper are [34] and [17]. While Tang *et al.* [34] proposed a SPL strategy in the context of dictionary learning, Jiang *et al.* [17] introduced the first hybrid full curriculum and SPL approach. However, none of them is able to handle data arising from multiple domains. Our approach does not only overcome this issue, but more importantly, is naturally able to incorporate domain specific partial curricula, thus benefiting from the wealth of information present in both domains. Moreover, opposite to [17] where the need to account for the difficulty of each training sample greatly limits the scalability of the method, in ACDL the adoption of partial curricula permits to tackle large scale problems.

3. ACADEMIC COUPLED DICTIONARY LEARNING FOR SBIR

As discussed in Section 1, in this paper we introduce the academic learning paradigm and instantiate it within the framework of coupled DL so to address SBIR. The proposed approach is articulated in two phases (see Figure 2). In the learning phase, the features extracted from the given set of images and corresponding sketches are fed into the ACDL method that learns the image and sketch-specific dictionaries together with the associated codes, taking into account the modality-specific partial curricula. At retrieval time, the representation of a sketch is computed by projecting its feature vector to the learned code. Image retrieval is performed by searching for the closet image code (in the Euclidean sense) to the given sketch code. In the following we present the proposed ACDL method (Section 3.1), its solver (Section 3.2), along with the technique we use for generating curricula (Section 3.3).

3.1 Academic Coupled Dictionary Learning

Let us assume the existence of K sketches and denote the features extracted from the k-th sketch as $\mathbf{f}^{\mathrm{S}}_k \in \mathbb{R}^{d_{\mathrm{S}}}$. Similarly, we assume the existence of L images and denote the features extracted from the l-th image as $\mathbf{f}^{\mathrm{I}}_l \in \mathbb{R}^{d_{\mathrm{I}}}$. We also assume two N-word dictionaries, one per modality: $\mathbf{Q}^{\mathrm{S}} = [\mathbf{q}^{\mathrm{S}}_n]_{n=1}^N \in \mathbb{R}^{d_{\mathrm{S}}\times N}$ and $\mathbf{Q}^{\mathrm{I}} = [\mathbf{q}^{\mathrm{I}}_n]_{n=1}^N \in \mathbb{R}^{d_{\mathrm{I}}\times N}$. Each sketch (resp. image) corresponds to a sparse code $\mathbf{c}^{\mathrm{S}}_k \in \mathbb{R}^N$ (resp. $\mathbf{c}^{\mathrm{I}}_l \in \mathbb{R}^N$). We also define $\mathbf{F}^{\mathrm{S}} = [\mathbf{f}^{\mathrm{S}}_1, \dots, \mathbf{f}^{\mathrm{S}}_K] \in \mathbb{R}^{d_{\mathrm{S}}\times K}$ as the matrix of all sketch features, and $\mathbf{F}^{\mathrm{I}}$, $\mathbf{C}^{\mathrm{S}}$ and $\mathbf{C}^{\mathrm{I}}$ analogously. Given the feature matrices $\mathbf{F}^{\mathrm{S}}$ and $\mathbf{F}^{\mathrm{I}}$, we propose to compute the associated dictionaries and codes by minimizing the following dictionary representation (DR) loss:

$$\begin{aligned}\min_{\mathbf{Q}^{\mathrm{S}},\mathbf{Q}^{\mathrm{I}},\mathbf{C}^{\mathrm{S}},\mathbf{C}^{\mathrm{I}}} \ell_{\mathrm{DR}} = \quad & \|\mathbf{F}^{\mathrm{S}} - \mathbf{Q}^{\mathrm{S}}\mathbf{C}^{\mathrm{S}}\|_{\mathcal{F}}^2 + \|\mathbf{F}^{\mathrm{I}} - \mathbf{Q}^{\mathrm{I}}\mathbf{C}^{\mathrm{I}}\|_{\mathcal{F}}^2 \\ & +\alpha\left(\|\mathbf{C}^{\mathrm{S}}\|_1 + \|\mathbf{C}^{\mathrm{I}}\|_1\right), \\ \text{s.t.} \quad & \|\mathbf{q}^{\mathrm{S}}_n\|, \|\mathbf{q}^{\mathrm{I}}_n\| \le 1 \quad \forall n, \end{aligned} \tag{1}$$

where $\alpha \geq 0$ is a regularization parameter and $\|\cdot\|_{\mathcal{F}}$ denotes the Frobenius norm. The constraints remove any scale ambiguities due to the matrix products $\mathbf{Q}^{\mathrm{S}}\mathbf{C}^{\mathrm{S}}$ and $\mathbf{Q}^{\mathrm{I}}\mathbf{C}^{\mathrm{I}}$, while the regularization terms induce sparsity in the learned codes.

We also assume a relational link between sketches and images in the training set. Ideally, each sketch corresponds to at least an image (*e.g.*, for sketch to photo face recognition [37] in the context of security and biometrics applications). Alternatively, the association among sketches and images is derived from image class information [12]. Generally speaking, in this paper we consider both intra-modality and cross-modality relationships, modeled by a non-negative weight matrix $\mathbf{W} \in \mathbb{R}^{+(K+L)\times(K+L)}$. Intuitively, the larger w_{pq} is, the stronger the relationship between the p-th and q-th codes is. Importantly, when $1 \leq p, q \leq K$ (respectively, $K < p, q \leq K + L$), w_{pq} relates two sketches (respectively, two images) creating an intra-modality link, otherwise w_{pq} relates a sketch and an image (cross-modality link). Interpreting $\mathbf{W}$ as the weight matrix of a graph and denoting the associated Laplacian matrix[1] by $\mathbf{L}$, the code-coupling (CC) loss is defined as:

$$\ell_{\mathrm{CC}} = \beta\,\mathrm{Tr}\left(\mathbf{C}^{\mathrm{J}}\mathbf{L}\mathbf{C}^{\mathrm{J}\top}\right) = \frac{1}{2}\beta\sum_{p,q=1}^{K+L} w_{pq}\|\mathbf{c}^{\mathrm{J}}_p - \mathbf{c}^{\mathrm{J}}_q\|^2, \tag{2}$$

where $\mathbf{C}^{\mathrm{J}} = [\mathbf{c}^{\mathrm{J}}_p]_{p=1}^{K+L} = [\mathbf{C}^{\mathrm{S}}\,\mathbf{C}^{\mathrm{I}}] \in \mathbb{R}^{N\times(K+L)}$ is the joint code matrix, and $\beta \geq 0$ is a regularization parameter controlling the importance of the relational knowledge.

As already mentioned in the introduction, and inspired by previous works [18, 16], we embrace the philosophy of SP learning in which there is a pacing binary variable $v^{\mathrm{S}}_k \in \{0, 1\}$ (respectively $v^{\mathrm{I}}_l \in \{0, 1\}$) associated to sketch k (respectively to image l), indicating whether the observation (sketch or image) has to be used for learning or not. Importantly, v^{S}_k and v^{I}_l are not fixed and evolve during the training phase. Additionally, we relax the binary condition and v^{S}_k and v^{I}_l become real numbers $v^{\mathrm{S}}_k, v^{\mathrm{I}}_l \in [0, 1]$ since this strategy has proven to be successful in several applications [45, 17]. By defining $\mathbf{V}^{\mathrm{S}} = \mathrm{diag}_k(v^{\mathrm{S}}_k) \in \mathbb{R}^{K\times K}$ as a diagonal matrix with entries $v^{\mathrm{S}}_1, \dots, v^{\mathrm{S}}_K$, $\mathbf{V}^{\mathrm{I}}$ analogously to $\mathbf{V}^{\mathrm{S}}$, and $\mathbf{V}^{\mathrm{J}} = \mathrm{diag}(\mathbf{V}^{\mathrm{S}}, \mathbf{V}^{\mathrm{I}})$,

[1] The Laplacian matrix of a graph with weight matrix $\mathbf{W}$ is defined as $\mathbf{L} = \mathbf{D} - \mathbf{W}$, where $\mathbf{D}$ is a diagonal matrix with $d_{pp} = \sum_q w_{pq}$.

we rewrite ℓ_{DR} and ℓ_{CC}:

$$\begin{aligned}\ell_{\text{DR}}(\mathbf{Q}^{\mathrm{S}},\mathbf{Q}^{\mathrm{I}},\mathbf{C}^{\mathrm{J}},\mathbf{V}^{\mathrm{J}}) &:= \|(\mathbf{F}^{\mathrm{S}}-\mathbf{Q}^{\mathrm{S}}\mathbf{C}^{\mathrm{S}})\mathbf{V}^{\mathrm{S}}\|_{\mathcal{F}}^2 \\ &+ \|(\mathbf{F}^{\mathrm{I}}-\mathbf{Q}^{\mathrm{I}}\mathbf{C}^{\mathrm{I}})\mathbf{V}^{\mathrm{I}}\|_{\mathcal{F}}^2 + \alpha\left(\|\mathbf{C}^{\mathrm{S}}\|_1+\|\mathbf{C}^{\mathrm{I}}\|_1\right), \qquad (3)\\ \ell_{\text{CC}}(\mathbf{C}^{\mathrm{J}},\mathbf{V}^{\mathrm{J}}) &:= \beta\,\mathrm{Tr}\left(\mathbf{C}^{\mathrm{J}}\mathbf{V}^{\mathrm{J}}\mathbf{L}\mathbf{V}^{\mathrm{J}\top}\mathbf{C}^{\mathrm{J}\top}\right). \qquad (4)\end{aligned}$$

Classically, and in order to avoid trivial solutions (*i.e.*, all v's set to zero) a self-paced regularizer is used:

$$r_{\text{SP}}(\mathbf{V}^{\mathrm{J}}) = \nu\left(\frac{1}{2}\sum_{k=1}^{K}\gamma(v_k^{\mathrm{S}}) + \frac{1}{2}\sum_{l=1}^{L}\gamma(v_l^{\mathrm{I}})\right) \qquad (5)$$

where $\gamma(v) = v^2 - 2v$, as in [45], and $\nu \geq 0$ is the self-paced parameter that increases at each iteration as in traditional SPL methods [18]. By further considering the code-coupling loss and the self-paced regularizer, the optimization problem (1) rewrites:

$$\begin{aligned}\min_{\mathbf{Q}^{\mathrm{S}},\mathbf{Q}^{\mathrm{I}},\mathbf{C}^{\mathrm{J}},\mathbf{V}^{\mathrm{J}}}\ & \ell_{\text{DR}}(\mathbf{Q}^{\mathrm{S}},\mathbf{Q}^{\mathrm{I}},\mathbf{C}^{\mathrm{J}},\mathbf{V}^{\mathrm{J}}) + \ell_{\text{CC}}(\mathbf{C}^{\mathrm{J}},\mathbf{V}^{\mathrm{J}}) - r_{\text{SP}}(\mathbf{V}^{\mathrm{J}})\\ \text{s.t. } & \|\mathbf{q}_n^{\mathrm{S}}\|, \|\mathbf{q}_n^{\mathrm{I}}\| \leq 1 \quad \forall n,\\ & 0 \leq v_k^{\mathrm{S}}, v_l^{\mathrm{I}} \leq 1 \quad \forall k,l\end{aligned}$$

The major methodological contribution of our work is to include per-modality partial curricula into the framework of DL with relational knowledge and to study its behavior within the SPL strategy already discussed. Subsequently, we assume the existence of two per-modality sets of constraints $\mathcal{C}^{\mathrm{S}}$ and $\mathcal{C}^{\mathrm{I}}$. Each element of the sets consists of an index pair meaning that if $(k,k') \in \mathcal{C}^{\mathrm{S}}$, then $v_k^{\mathrm{S}} < v_{k'}^{\mathrm{S}}$ and learning should be performed considering a priori $\mathbf{f}_{k'}^{\mathrm{S}}$ before $\mathbf{f}_k^{\mathrm{S}}$, as it corresponds to an easier sample. Depending on the way the curricula are constructed $\mathcal{C}^{\mathrm{S}}$ could contain incompatibilities, for instance, $\{(k,k'),(k',k''),(k'',k)\} \subset \mathcal{C}^{\mathrm{S}}$. In addition, the cross-modal terms could also induce incompatibilities between the two modalities. Therefore, it is desirable to relax the constraints using a set of slack variables $\xi_{kk'}^{\mathrm{S}}$ and $\xi_{ll'}^{\mathrm{I}}$. In all, the optimization problem for *academic coupled dictionary learning* writes:

$$\begin{aligned}\min_{\mathbf{Q}^{\mathrm{S}},\mathbf{Q}^{\mathrm{I}},\mathbf{C}^{\mathrm{J}},\mathbf{V}^{\mathrm{J}},\boldsymbol{\xi}^{\mathrm{J}}}\ & \ell_{\text{DR}}(\mathbf{Q}^{\mathrm{S}},\mathbf{Q}^{\mathrm{I}},\mathbf{C}^{\mathrm{J}},\mathbf{V}^{\mathrm{J}}) + \ell_{\text{CC}}(\mathbf{C}^{\mathrm{J}},\mathbf{V}^{\mathrm{J}})\\ & - r_{\text{SP}}(\mathbf{V}^{\mathrm{J}}) + r_{\text{PC}}(\boldsymbol{\xi}^{\mathrm{J}}) \qquad (6)\\ \text{s.t. } & \|\mathbf{q}_n^{\mathrm{S}}\|, \|\mathbf{q}_n^{\mathrm{I}}\| \leq 1 \quad \forall n,\\ & 0 \leq v_k^{\mathrm{S}}, v_l^{\mathrm{I}} \leq 1 \quad \forall k,l\\ & v_k^{\mathrm{S}} - v_{k'}^{\mathrm{S}} < \xi_{kk'}^{\mathrm{S}},\ \xi_{kk'}^{\mathrm{S}} \geq 0,\ \forall (k,k') \in \mathcal{C}^{\mathrm{S}}\\ & v_l^{\mathrm{I}} - v_{l'}^{\mathrm{I}} < \xi_{ll'}^{\mathrm{I}},\ \xi_{ll'}^{\mathrm{I}} \geq 0,\ \forall (l,l') \in \mathcal{C}^{\mathrm{I}},\end{aligned}$$

where $\boldsymbol{\xi}^{\mathrm{J}} = [[\xi_{kk'}^{\mathrm{S}}]_{(k,k')\in\mathcal{C}^{\mathrm{S}}}[\xi_{ll'}^{\mathrm{I}}]_{(l,l')\in\mathcal{C}^{\mathrm{I}}}]$ is the vector of all slack variables and r_{PC} is the partial curricula regularizer regulated by the parameter $\gamma \geq 0$ and defined as:

$$r_{\text{PC}}(\boldsymbol{\xi}^{\mathrm{J}}) = \gamma\left(\sum_{(k,k')\in\mathcal{C}^{\mathrm{S}}}\xi_{kk'}^{\mathrm{S}} + \sum_{(l,l')\in\mathcal{C}^{\mathrm{I}}}\xi_{ll'}^{\mathrm{I}}\right). \qquad (7)$$

The algorithm we propose to solve (6) is outlined in Algorithm 1 and detailed in the following section.

3.2 Solving ACDL

The ACDL optimization problem is not jointly convex in all variables. However, efficient alternate optimization techniques can solve it since it is convex on $\{\mathbf{Q}^{\mathrm{S}},\mathbf{Q}^{\mathrm{I}}\}$, $\{\mathbf{C}^{\mathrm{J}}\}$ and $\{\mathbf{V}^{\mathrm{J}},\boldsymbol{\xi}^{\mathrm{J}}\}$ when the other two sets of variables are fixed.

Algorithm 1 ACDL optimisation procedure.

```
Input: $\mathbf{F}^{\mathrm{S}}$, $\mathbf{F}^{\mathrm{I}}$ and the parameters $\alpha$, $\beta$, $\gamma$.
Initialize $\mathbf{Q}^{\mathrm{S}}$, $\mathbf{C}^{\mathrm{S}}$, $\mathbf{Q}^{\mathrm{I}}$, $\mathbf{C}^{\mathrm{I}}$ as described in Section 4.
repeat
    Update $\mathbf{V}^{\mathrm{J}}$ and $\boldsymbol{\xi}^{\mathrm{J}}$ following (12);
    Update $\mathbf{C}^{\mathrm{S}}$, $\mathbf{C}^{\mathrm{I}}$ with (10);
    Update $\mathbf{Q}^{\mathrm{S}}$, $\mathbf{Q}^{\mathrm{I}}$ by solving (8);
    Increase $\nu$ so to enlarge the training set;
until All the training data points are selected.
Output: $\mathbf{Q}^{\mathrm{S}}$, $\mathbf{C}^{\mathrm{S}}$, $\mathbf{Q}^{\mathrm{I}}$, $\mathbf{C}^{\mathrm{I}}$.
```

Solve for $\mathbf{Q}^{\mathrm{S}}$ and $\mathbf{Q}^{\mathrm{I}}$. Fixing $\mathbf{C}^{\mathrm{J}}$, $\mathbf{V}^{\mathrm{J}}$ and $\boldsymbol{\xi}^{\mathrm{J}}$, the optimization problem for $\mathbf{Q}^{\mathrm{S}}$ (analogously for $\mathbf{Q}^{\mathrm{I}}$) writes:

$$\min_{\mathbf{Q}^{\mathrm{S}}} \left\|(\mathbf{F}^{\mathrm{S}}-\mathbf{Q}^{\mathrm{S}}\mathbf{C}^{\mathrm{S}})\mathbf{V}^{\mathrm{S}}\right\|_2^2 \quad \text{s.t. } \left\|\mathbf{q}_k^{\mathrm{S}}\right\| \leq 1. \qquad (8)$$

This problem is a Quadratically Constrained Quadratic Program (QCQP) that can be solved using gradient descent with e.g. Lagrangian duality [19].

Solve for $\mathbf{C}^{\mathrm{J}}$. By fixing $\mathbf{Q}^{\mathrm{S}}$, $\mathbf{Q}^{\mathrm{I}}$, $\mathbf{V}^{\mathrm{J}}$ and $\boldsymbol{\xi}^{\mathrm{J}}$ the optimization function for the codes can be rewritten as:

$$\begin{aligned}f(\mathbf{C}^{\mathrm{J}}) = &\left\|\left(\mathbf{F}^{\mathrm{S}}-\mathbf{Q}^{\mathrm{S}}\mathbf{C}^{\mathrm{S}}\right)\mathbf{V}^{\mathrm{S}}\right\|_{\mathcal{F}}^2 + \left\|\left(\mathbf{F}^{\mathrm{I}}-\mathbf{Q}^{\mathrm{I}}\mathbf{C}^{\mathrm{I}}\right)\mathbf{V}^{\mathrm{I}}\right\|_{\mathcal{F}}^2\\ &+ \alpha\left\|\mathbf{C}^{\mathrm{J}}\right\|_1 + \beta\,\mathrm{Tr}\left(\mathbf{C}^{\mathrm{J}}\mathbf{V}^{\mathrm{J}}\mathbf{L}\mathbf{V}^{\mathrm{J}\top}\mathbf{C}^{\mathrm{J}\top}\right). \qquad (9)\end{aligned}$$

According to FISTA [2], f can be viewed as a proximal regularization problem, solved using the following recursion (over r):

$$\mathbf{C}_r^{\mathrm{J}} = \underset{\mathbf{C}^{\mathrm{J}}}{\operatorname{argmin}}\left\{\frac{\|\mathbf{C}^{\mathrm{J}}-\mathbf{C}_{r-1}^{\mathrm{J}}+t_r\nabla f(\mathbf{C}_{r-1}^{\mathrm{J}})\|_{\mathcal{F}}^2}{2t_r} + \alpha\left\|\mathbf{C}^{\mathrm{J}}\right\|_1\right\}, \qquad (10)$$

where $t_r > 0$ is the step size and $\nabla f(\mathbf{C}^{\mathrm{J}}) = [\nabla f(\mathbf{C}^{\mathrm{S}})\ \nabla f(\mathbf{C}^{\mathrm{I}})]$ is the concatenation of the two gradients defined as:

$$\begin{aligned}\nabla f(\mathbf{C}^{\mathrm{S}}) = &\ 2\mathbf{Q}^{\mathrm{S}\top}(\mathbf{Q}^{\mathrm{S}}\mathbf{C}^{\mathrm{S}}-\mathbf{F}^{\mathrm{S}})(\mathbf{V}^{\mathrm{S}})^2\\ &+ 2\beta\left(\mathbf{C}^{\mathrm{S}}\mathbf{V}^{\mathrm{S}}\mathbf{L}^{\mathrm{S}} + \mathbf{C}^{\mathrm{I}}\mathbf{V}^{\mathrm{I}}\mathbf{L}^{\mathrm{IS}}\right)\mathbf{V}^{\mathrm{S}}, \qquad (11)\end{aligned}$$

where the sublaplacian matrices are taken from the Laplacian matrix as $\mathbf{L} = [\mathbf{L}^{\mathrm{S}}\,\mathbf{L}^{\mathrm{SI}};\mathbf{L}^{\mathrm{IS}}\,\mathbf{L}^{\mathrm{I}}]$. The second gradient, $\nabla f(\mathbf{C}^{\mathrm{I}})$ is defined analogously to $\nabla f(\mathbf{C}^{\mathrm{S}})$. Moreover, (10) is a standard LASSO problem whose optimal solution can be found using the feature-sign search algorithm in [19].

Solve for $\mathbf{V}^{\mathrm{J}}$ and $\boldsymbol{\xi}^{\mathrm{J}}$. By fixing the dictionaries $\mathbf{Q}^{\mathrm{S}}$, $\mathbf{Q}^{\mathrm{I}}$ and the codes $\mathbf{C}^{\mathrm{J}}$, we can rewrite the joint optimization problem with respect to $\mathbf{V}^{\mathrm{J}}$ and $\boldsymbol{\xi}^{\mathrm{J}}$ as a quadratic programming problem with a set of linear inequality constraints. Denoting the optimization variable as $\mathbf{y} = [[v_k^{\mathrm{S}}]_k[v_l^{\mathrm{I}}]_l[\xi_{kk'}^{\mathrm{S}}]_{kk'}[\xi_{ll'}^{\mathrm{I}}]_{ll'}] \in \mathbb{R}^{K+L+C^{\mathrm{S}}+C^{\mathrm{I}}}$, the problem writes:

$$\begin{aligned}\min_{\mathbf{y}}\ & \mathbf{y}^\top\mathbf{R}\mathbf{y} + \mathbf{b}^\top\mathbf{y} \qquad (12)\\ \text{s.t. } & \mathbf{G}\mathbf{y} \leq \mathbf{h},\end{aligned}$$

where the values of $\mathbf{R}$, $\mathbf{b}$, $\mathbf{G}$ and $\mathbf{h}$ are defined in the following. $\mathbf{R}$ is a $(K+L+C^{\mathrm{S}}+C^{\mathrm{I}}) \times (K+L+C^{\mathrm{S}}+C^{\mathrm{I}})$ matrix with all zeros except for the first $(K+L) \times (K+L)$ block. More precisely:

$$R_{pq} = \begin{cases}\|\mathbf{f}_p^{\mathrm{S}} - \mathbf{Q}^{\mathrm{S}}\mathbf{c}_p^{\mathrm{S}}\|^2 - \nu/2 & q = p \leq K\\ \|\mathbf{f}_{p-K}^{\mathrm{I}} - \mathbf{Q}^{\mathrm{I}}\mathbf{c}_{p-K}^{\mathrm{I}}\|^2 - \nu/2 & K < q = p \leq L+K\\ \beta w_{pq}\|\mathbf{c}_p^{\mathrm{J}} - \mathbf{c}_q^{\mathrm{J}}\|^2 & 1 \leq p \neq q \leq K+L\\ 0 & \text{otherwise}\end{cases}$$

and $\mathbf{b} = [\nu\mathbf{1}_{K+L}^\top\ \gamma\mathbf{1}_{C^{\mathrm{S}}+C^{\mathrm{I}}}^\top]^\top$, where $\mathbf{1}_K$ is a $K \times 1$ vector filled with ones. $\mathbf{G}$ and $\mathbf{h}$ represent the inequality and bound constraints in (6) and their derivation is straightforward. Since there are $2(K + L + C^{\mathrm{s}} + C^{\mathrm{I}})$ constraints, $\mathbf{G} \in \mathbb{R}^{2(K+L+C^{\mathrm{s}}+C^{\mathrm{I}})\times(K+L+C^{\mathrm{s}}+C^{\mathrm{I}})}$ and $\mathbf{h} \in \mathbb{R}^{2(K+L)+C^{\mathrm{s}}+C^{\mathrm{I}}}$.

3.3 Laplacian and Curricula Construction

In this section we describe how we construct the modality-specific curricula and the Laplacian matrix representing the relational knowledge. However, it is worth noting that ACDL is a general framework and other design choices are possible. We build both the curricula and the Laplacian in the training set from the sketch and image features and a group association, that could arise from the class membership or from unsupervised clustering. In our experiments, we also devised a protocol to construct a curriculum for sketches from human manual annotations.

3.3.1 Laplacian construction

To build the Laplacian matrix (computed from the weights w_{pq}), the intra-modality relationships are defined using the Gaussian kernel and the inter-modality with group association, as in [44]:

$$w_{pq} = \begin{cases} \mathrm{e}^{-\|\mathbf{f}^{\mathrm{s}}_p - \mathbf{f}^{\mathrm{s}}_q\|_2^2/2\sigma^2}, & p, q \le K \\ \mathrm{e}^{-\|\mathbf{f}^{\mathrm{I}}_{p-K} - \mathbf{f}^{\mathrm{I}}_{q-K}\|_2^2/2\sigma^2}, & K < p, q \\ 1, & \begin{array}{l} p \le K < q \text{ and } p \sim q \\ q \le K < p \text{ and } q \sim p \end{array} \\ 0, & \text{otherwise,} \end{cases} \tag{13}$$

where σ is the Gaussian kernel parameter fixed to 1 with no significant performance variation around this value. The symbol $\sim$ indicates samples belonging to the same cluster/class.

3.3.2 Curriculum construction

Regarding the curricula construction, as stated above, a fundamental aspect of the the proposed ACDL framework is the possibility to handle partial curricula. Previous CL or hybrid CL-SPL methods [3, 17] instead assume that a full curriculum, *i.e.* a complete order of samples, is provided. This is a strong assumption that may not be satisfied in real-world large-scale tasks. On the one hand, even if automatic measures of the easiness of an image [20] have been developed, these metrics are accurate up to some extent and therefore deriving a full ranking from these measures may be inappropriate. On the other hand, manually annotating the entire set of images represents a huge human workload, highly demanding for medium-scale problems and chimeric for large-scale datasets. In addition, if the multimedia dataset is gathered incrementally, the cost of updating the curriculum grows with the size of the dataset.

The partial curricula for the image domain is obtained by means of an automated procedure based on previous studies [20, 1]. Intuitively, easy images are those containing non-occluded high-resolution objects in low-cluttered background. Previous works [20] proposed to define the easiness of an image from the "objectness" measures [1]. In the same line of though, we compute the easiness measure as the median of the 30 highest "objectness" scores among a set of 1,000 window proposals. This procedure approximates the easiness of a training image. Notice that, two images with largely different scores are likely to correspond to samples with different easiness. On the contrary, if the scores are similar, imposing that the image with the lowest score is the easiest in the pair may induce some errors. The constraint associated to an image pair is included in $\mathcal{C}^{\mathrm{I}}$ only if the difference of their associated scores exceeds a certain threshold δ^{I} (*i.e.* if one of the images in the pair is significantly easier than the other).

Contrary to the image domain, there is no widely-accepted procedure to define the easiness of a sketch. Therefore, we consider two methods for constructing the partial curriculum in the sketch domain. First, an automatic method that follows again the philosophy of [1]. Given a sketch, we randomly sample 100 windows at different scales and positions. For each window we compute the "edgeness" score, representing the edge density within the window as proposed in [1]. Intuitively, the edgeness should implement the rationale that easy sketches are those with more details. As previously done for images, the constraint associated to a pair of sketches is included into $\mathcal{C}^{\mathrm{s}}$ only if their measure of edgeness differs by at least δ^{s}. Second, a semi-automatic strategy for building the partial curricula of the sketches by including human annotators in the loop. A naive retrieval method based on SHOG features [9] generates potential constraints (pairs of sketches). In details, we pair each sketch with the closest among the cluster/class. The human annotator is then queried which sketch is easier to learn from. Ten PhD students (6 male, 4 female) of age 24.3 ± 1.4 (mean, standard deviation) performed the annotation after being instructed that "easy" sketches meant sketches with more details. Importantly, since ACDL is specifically designed to handle partial curricula, annotators had the possibility to "skip" sketch pairs if they were unable to decide. A simple GUI (Fig.3) was developed for annotation.

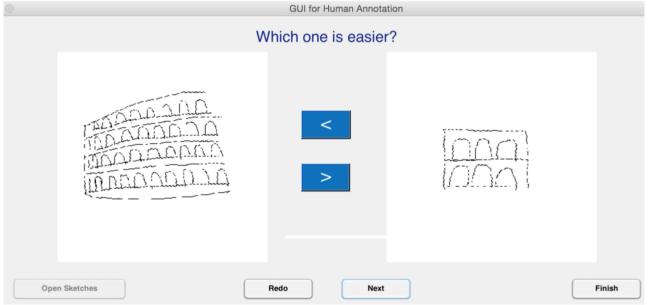

Figure 3: The graphical interface used for annotation. Easy sketches are those with more details and easy images are those with non-occluded high-resolution objects in low-cluttered background.

4. EXPERIMENTS

To evaluate the effectiveness of our approach, we conduct extensive experiments on three publicly available datasets: the CUHK Face Sketch (CUFS) [37], the Flickr15k [12] and the Queen-Mary SBIR [21]. In our experiments, the dictionaries of ACDL were initialized with joint DL [42] when both features have the same dimension and with modality-independent DL otherwise.

4.1 Experiments on CUFS

The CUFS dataset is a very popular benchmark for sketch-to-face-photo recognition. CUFS contains sketch-photo image pairs of 188 CUHK students. The recognition task is to identify the face photo corresponding to a given sketch. We evaluate the performance of ACDL on CUFS and compare it to other cross-domain retrieval methods and previous coupled DL approaches.

4.1.1 Settings

In our experiments, 88 sketch-photo image pairs are randomly selected for training the model, and the remaining 100 pairs are used for testing, as in [33]. To facilitate comparison with previous works [42, 13], we use raw pixels as features. We compare the proposed ACDL method to seven baseline methods: canonical correlation analysis (CCA), partial least squares (PLS) [30], bilinear model (Bil) [35], semi-coupled dictionary learning (SCDL) [36], joint dictionary learning (JDL) [42] and coupled dictionary learn-

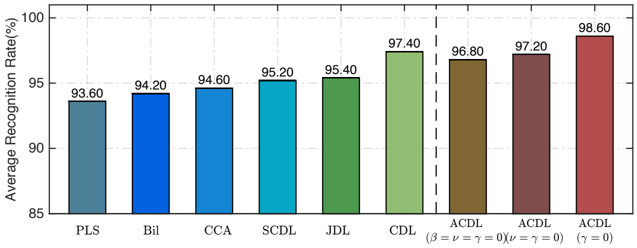

Figure 4: Average recognition rate for all methods on CUFS.

ing (CDL) [13]. For the bilinear model, we used 70 PLS bases and 50 eigenvectors (see [30]). ACDL, SCDL, JDL and CDL are all DL-based approaches and we set the dictionary size to 50. In all cases, the recognition is performed using the nearest neighbor on the newly learned sparse representation as in [30, 13]. For ACDL, we explicitly evaluate the importance of the relational knowledge (β) and of self-pacing (ν). Instead, as both images and sketches are quite homogeneous (*i.e.*, sketches were drawn by experts, faces in images are centered and equally illuminated), we did not use any curriculum ($\gamma = 0$), The parameters α, β were set by cross-validation to 1 and 5, respectively.

4.1.2 Results

Figure 4 reports average recognition results over five trials. ACDL($\gamma = 0$) achieves the best average recognition rate: 98.6%. Remarkably, ACDL($\gamma = 0$) outperforms CDL, which is the best of the DL-based approaches. Importantly, we notice that the effect of the relational knowledge is crucial in the performance of the overall method (CDL also uses relational knowledge). Among the compared methods, SCDL, JDL and CDL are the strongest competitors, achieving 95.2%, 95.4% and 97.4% recognition rate respectively. This means that DL for cross-modal retrieval is an effective strategy. We also remark that ACDL($\gamma = 0$) outperforms the other two versions of ACDL, suggesting that the relational knowledge within the SP learning framework is beneficial for accurate retrieval.

4.2 Experiments on Flickr15K

Flickr15k dataset is a widely used dataset for SBIR, containing around 14,660 images collected from Flickr and 330 free-hand sketches drawn by 10 non-expert sketchers. All samples are classified into 33 categories with a manual annotation for each sample. As this dataset does not provide a training set, to evaluate our approach, we partition the dataset into a training set with randomly chosen 40% samples and a test set with the remaining samples.[2]

4.2.1 Settings

We compare the retrieval performance of ACDL against several state of the art SBIR baselines, including SHOG [9] and SIFT, SSIM, GFHOG tested in [12], Structure Tensor [8], Learned Key Shapes (LKS) [27] and PerceptualEdge [26]. The first five baselines extract low-level features (SHOG, SIFT, SSIM, GFHOG and StructureTensor) over Canny edge maps (for images) and sketched images, and then use a bag-of-words (BOW) approach to obtain both image and sketch representations. LKS learns mid-level sketch patterns named keyshapes. The learned keyshapes are used to construct image and sketch descriptors. PerceptualEdge uses an edge grouping framework to create synthesized sketches from images. The retrieval is performed by querying the synthesized sketches instead of the images directly. For LKS and PerceptualEdge, we use the original codes provided by the authors with the same parameter setting described in the associated papers, and we reimplemented other baselines whose codes are not publicly available. All the methods are evaluated on the same test set for a fair comparison.

Table 1: Flickr15K dataset: comparison of different methods.

Methods	mAP
StructureTensor [8]	0.0801
SIFT [12]	0.0967
SSIM [12]	0.1068
SHOG [9]	0.1152
GFHOG [12]	0.1245
LKS [27]	0.1640
PerceptualEdge [26]	0.1741
ACDL$_{\mathrm{GFHOG}}$	0.1675
ACDL$_{\mathrm{LKS}}$ ($\gamma = \nu = 0$)	0.2278
ACDL$_{\mathrm{LKS}}$	0.2495
ACDL$_{\mathrm{CNN+LKS}}$	0.2656

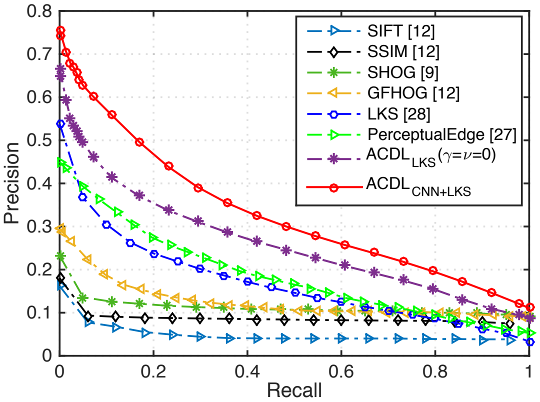

Figure 5: Precision recall curves of the different methods on Flirckr15K dataset.

We evaluate the proposed ACDL with different settings, including (i) ACDL$_{\mathrm{GFHOG}}$: ACDL using GFHOG features for both image and sketch domains; (ii) ACDL$_{\mathrm{LKS}}$, ($\gamma = \nu = 0$): ACDL without the curricula and self-pacing, using LKS features for both image and sketch domains; (iii) ACDL$_{\mathrm{LKS}}$: ACDL using LKS features for both image and sketch domains; (iv) ACDL$_{\mathrm{CNN+LKS}}$: ACDL using CNN features for image domain (*i.e.*, features extracted from the sixth layer of the Caffe reference network trained on ImageNet [14]) and LKS features for sketch domain. The sketch curriculum, when used, is constructed using 60% of human annotations, since we did not observe any significant differences between the automatic and the manual procedures (see Section 4.4). For all ACDL methods, we set α, β, γ with cross-validation and $N = 1000$.

4.2.2 Results

The quantitative SBIR performance comparison is shown in Table 1, reporting the mean average precision (mAP), and in Figure 5, depicting the precision-recall (PR) curve. We observe that ACDL$_{\mathrm{CNN+LKS}}$ obtains the best mAP of 0.2656, which corresponds to a significant performance increase (7.13 points improvement) with respect to the best state-of-the-art method (0.1741 of PerceptualEdge [26]). The advantage of the academic learning paradigm and its instantiation under the framework of dictionary

[2] Due to the random training/test splitting, performance may differ from results reported in previous studies. Our experimental protocol will be made publicly available.

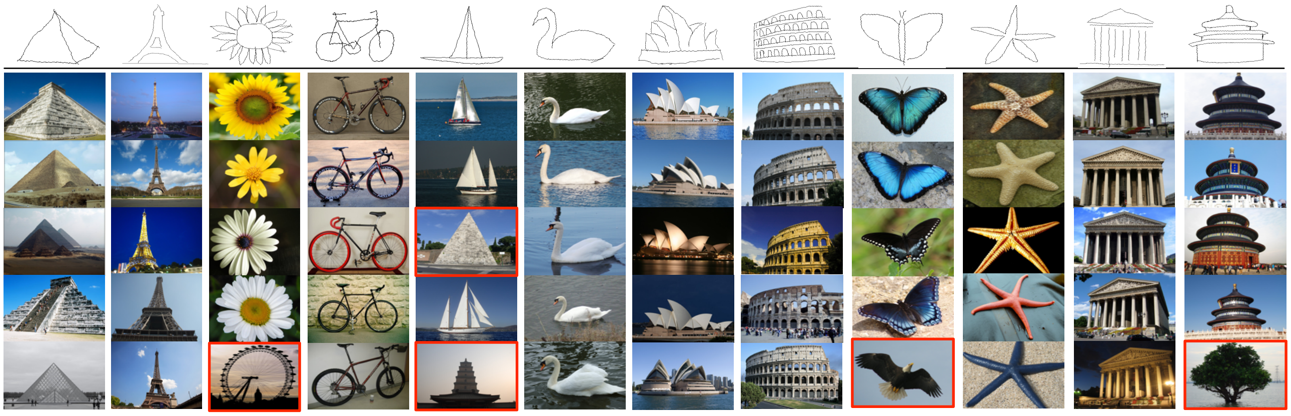

Figure 6: Top 5 retrieval results with sample query sketches in the Flickr15K dataset. Red boxes show false positive retrievals.

Table 2: Benchmark on the Queen-Mary SBIR dataset.

Methods	mAP
SIFT [12]	0.0685
SSIM [12]	0.0745
SP-SHOG [8]	0.0804
SP-GFHOG [12]	0.0858
LKS [27]	0.1182
PerceptualEdge [26]	0.1246
$\mathrm{ACDL}_{\mathrm{LKS}}$ ($\gamma = \nu = 0$)	0.1265
$\mathrm{ACDL}_{\mathrm{LKS}}$	0.1467
$\mathrm{ACDL}_{\mathrm{CNN+LKS}}$	0.1598

learning is clear and independent of the features used. Indeed, $\mathrm{ACDL}_{\mathrm{GFHOG}}$ and $\mathrm{ACDL}_{\mathrm{LKS}}$ compares favorably to GFHOG [12] and to LKS [27], with 4.3 points and 7.61 points improvement, respectively. Similarly, the clear performance gap when $\mathrm{ACDL}_{\mathrm{LKS}}$ is compared to the $\mathrm{ACDL}_{\mathrm{LKS}}$, ($\gamma = \nu = 0$), demonstrates the usefulness of the combination of a variable pacing rate under the prior knowledge of multiple partial curricula. The fact that the best performance is obtained with $\mathrm{ACDL}_{\mathrm{CNN+LKS}}$ confirms our original intuition that different features can represent better the two different modalities. Finally, Figure 6 shows some qualitative results (top-five retrieved images) associated with the proposed method.

4.3 Experiments on Queen-Mary SBIR

The Queen-Mary SBIR dataset [21] contains 1,120 sketches and 7,267 images, built by intersecting 14 common categories from the Eitz 20,000 sketch dataset [7] and the PASCAL VOC dataset [10]. In this dataset, cluttered background conditions and significant scale variations greatly increase the complexity of the retrieval task. We use the given training and test set. While this dataset was conceived for fine-grained SBIR, since our task is category-level SBIR, we only use image-level category annotations.

4.3.1 Settings

We compare our approach to SIFT [12], SSIM [12], SP-SHOG [21], SP-GFHOG [12], LKS [27], PerceptualEdge [26]. SP-HOG and SP-GFHOG use a spatial pyramid model on SHOG and GFHOG features respectively to construct a new representation. The QueenMary SBIR dataset is designed for complex SBIR task, therefore we do not consider the weakest baselines (*i.e.*, StructureTensor) evaluated on Flickr15k dataset, and use SP-SHOG and SP-GFHOG instead of SHOG and GFHOG, as it has been demonstrated that the spatial pyramid model is able to obtain more robust image representations than BOW. We evaluate the proposed ACDL using the settings described in Section 4.2, *i.e.* $\mathrm{ACDL}_{\mathrm{LKS}}$ ($\gamma = \nu = 0$), $\mathrm{ACDL}_{\mathrm{LKS}}$ and $\mathrm{ACDL}_{\mathrm{CNN+LKS}}$, and the strategy for constructing the curricula. We set $N = 1000$ and cross-validate for the other parameters.

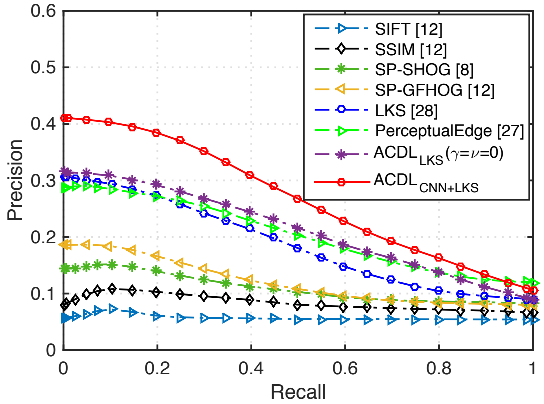

Figure 7: Precision recall curves of the different methods on Queen-Mary SBIR dataset.

4.3.2 Results

From Table 2, we observe that the mAP of $\mathrm{ACDL}_{\mathrm{CNN+LKS}}$ is 0.1598, *i.e.*, 3.52 points better than the best of all the baselines, which is not a trivial improvement on this challenging dataset. $\mathrm{ACDL}_{\mathrm{CNN+LKS}}$ also outperforms $\mathrm{ACDL}_{\mathrm{LKS}}$, clearly demonstrating the effectiveness of using different descriptors for sketches and images in SBIR. We also believe that LKS features are not robust enough to represent objects with various poses and cluttered backgrounds, as in Queen-Mary SBIR dataset. $\mathrm{ACDL}_{\mathrm{LKS}}$ obtains a clear improvement over $\mathrm{ACDL}_{\mathrm{LKS}}$, ($\mu = \gamma = 0$), further verifying the usefulness of the proposed learning scheme. Figure 7 depicts the PR curves of the different methods, confirming the findings so far. Additionally, Figure 8 shows the retrieval performance for each category-level task. We observe that for most of the classes (except for Airplane and Bicycle), $\mathrm{ACDL}_{\mathrm{CNN+LKS}}$ significantly outperforms all the baseline methods. Finally, Figure 9 reports the top 5 retrieval results of $\mathrm{ACDL}_{\mathrm{CNN+LKS}}$ for six sample query sketches.

4.4 In-depth analysis of ACDL

Finally, we provide some additional results and an empirical analysis (on the Flickr15k dataset) of the proposed approach regarding several aspects.

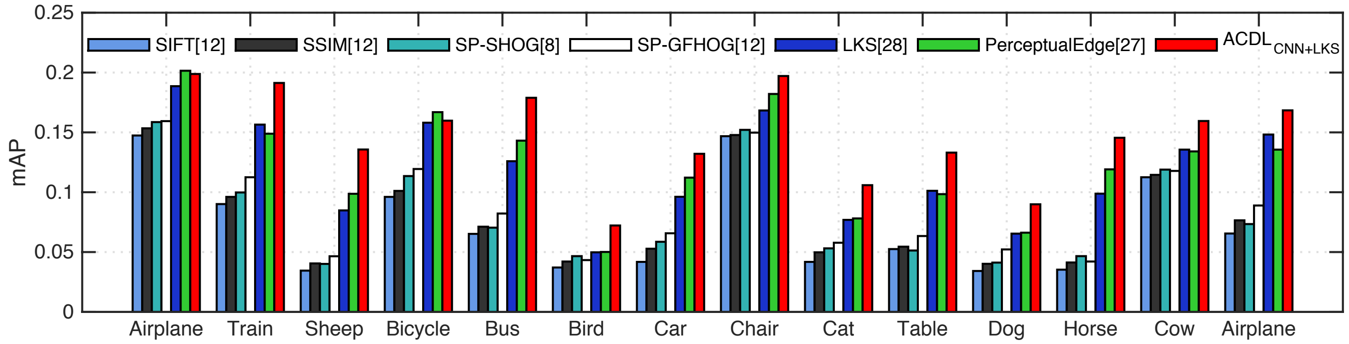

Figure 8: Retrieval performance comparison for each category on Queen-Mary SBIR dataset.

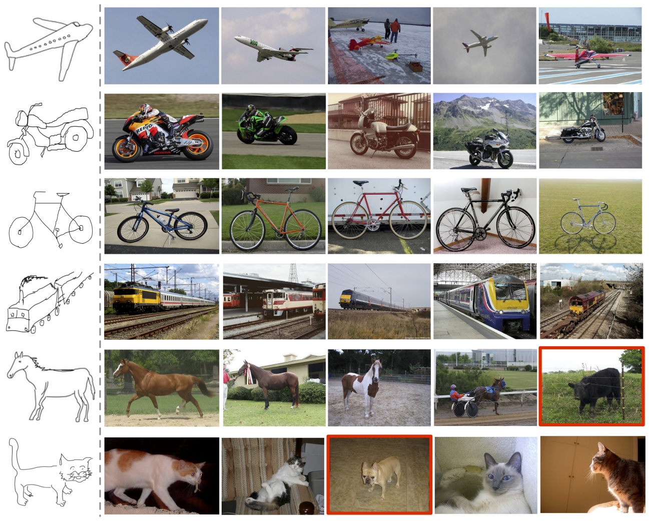

Figure 9: Top 5 retrieved images (2nd to 6th columns) using the query sketch samples (1st column) in the QueenMary SBIR dataset. Red boxes show false positive retrievals.

4.4.1 *Sensitivity analysis*

In order to assess the influence of the different model parameters on ACDL retrieval performance, we report a sensitivity analysis around the working point. Figure 10 shows the mAP as a function of α, β, γ, N. The analysis on α, β and γ is in the range $[10^{-2}, 10^{2}]$, on N in the range $[200, 2400]$. The plots clearly show that, while the method is sensitive to α and β, its performance does not change drastically with a wide range of γ and N. The sensitivity on β was already found in the experiments on CUFS and by previous researchers [13].

4.4.2 *Analysis of curriculum construction*

In our approach, the human and the automated annotator can both be employed to construct the partial curricula. To investigate the difference in terms of the final retrieval performance, we plot the mAP as a function of ρ, the proportion of constraints used for the sketch curriculum relative to the number of possible constraints. The proportion of image constraints was set to 10% of all possible constraints. Figure 11 shows these two plots. While the human annotations build a more useful partial curricula, the differences when compared with the automated curricula constructions are tiny. Interestingly, the excess of constraints leads to a slight decrease in performance, which is an experimental finding that supports the use of partial curricula for SBIR.

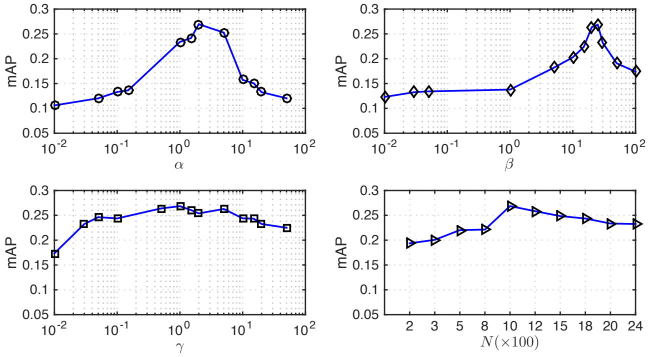

Figure 10: Empirical analysis of the model parameters: α, β, γ and the dictionary size N.

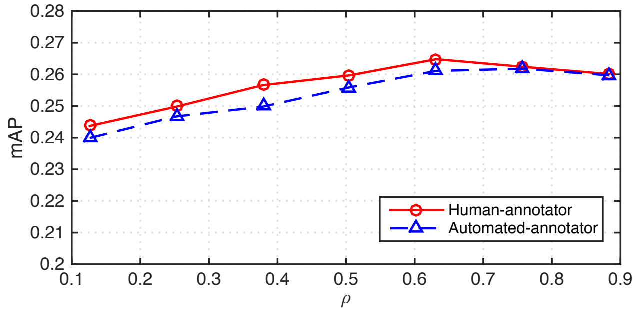

Figure 11: Performance of the automated and human-annotated sketch partial curricula.

4.4.3 *Convergence analysis*

Figure 12 plots the objective function value as a function of the iteration index for the proposed learning models ACDL and ACDL ($\gamma = \nu = 0$) for one of our experiments. These results clearly demonstrate the convergence of the proposed iterative optimization procedure. ACDL and ACDL ($\gamma = \nu = 0$) attain a stable solution within less than 40 iterations and 30 iterations respectively, proving the efficiency of the algorithm proposed to solve the ACDL optimization problem. It is worth noting that ACDL reaches a much lower local minima than ACDL ($\gamma = \nu = 0$) (6.8×10^4 vs. 1.98×10^5), verifying the beneficial effect of the proposed partial curriculum and self-paced learning scheme.

4.4.4 *Computational analysis*

To perform an empirical analysis of the computational overhead, we ran experiments on a PC with a quad core (2.1 GHz) CPU, 64GB RAM and an Nvidia Tesla K40 GPU. The proposed SBIR approach is implemented mostly in Matlab, and partially in C++

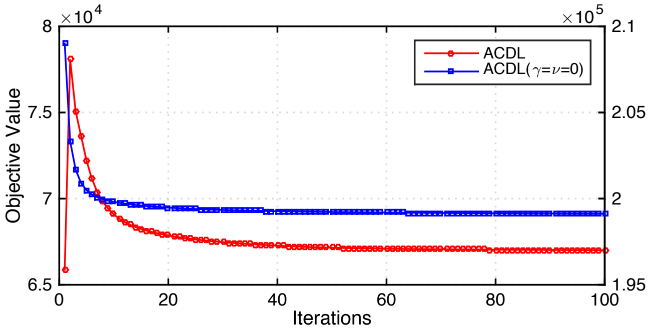

Figure 12: Convergence in terms of objective value: the left and right y axes show ACDL and ACDL($\gamma = \nu = 0$), respectively.

Table 3: Computational cost of the different training steps.

Phase	Component	Time overhead
Feature extraction	CNN (for images) LKS (for sketches)	0.04 sec/sample 1.1 sec/sample
Training	Curriculum Construction ACDL ACDL($\nu = \gamma = 0$)	8 min 27 min 21 min

(the most computationally expensive components). For LKS features [27], we use a C++ implementation, and wrap it to be called from Matlab. For CNN image features, the Caffe reference network pre-trained on ImageNet [14] is used to extract features from the sixth layer. In the following, we conduct the analysis from the off-line training phase and the on-line retrieval phase respectively.

Off-line retrieval time

The training phase of our method mainly contains three steps: (i) feature extraction, (ii) curriculum construction and (iii) ACDL optimization. Table 3 reports computational times of different components of the method. For the feature extraction, we consider CNN features from the image domain, which cost 0.04 seconds per image sample. The CNN image features are extracted with the GPU. LKS is used to extract features from the sketch domain, and obtains a speed of 1.1 seconds per sketch sample. The automated curriculum construction takes around 8 minutes, and training ACDL and ACDL ($\nu = \gamma = 0$) with 50 iterations cost 27 and 21 minutes, respectively.

Figure 13 shows elapsed times of updating $\mathbf{Q}_\text{I}$, $\mathbf{Q}_\text{S}$, $\mathbf{C}_\text{I}$ & $\mathbf{C}_\text{S}$ and $\mathbf{V}$ & ξ of each iteration within the model training. The input for ACDL is CNN features (for images) and LKS features (for sketches). 4833 and 132 are the number of training image and sketch samples, respectively. For curricula, we use the same strategy as in Section 4.2 (10% of image constraints and 60% of sketch constraints) as at this ratio the retrieval performance and the computational overhead reach good balance. The average elapsed times of these four components within the first 50 iterations are 6.5, 4.8, 12.1, 6.9 seconds, respectively. Updating $\mathbf{V}$ & ξ occupies only around 20% of the entire training time, demonstrating the reasonable computational efficiency of the proposed ACDL scheme.

On-line retrieval time

Online retrieval efficiency is a very important performance index for SBIR, especially for large-scale retrieval scenarios. We compare the online retrieval time of our method with the state-of-the-art, and the results are shown in Figure 14. Our method (here

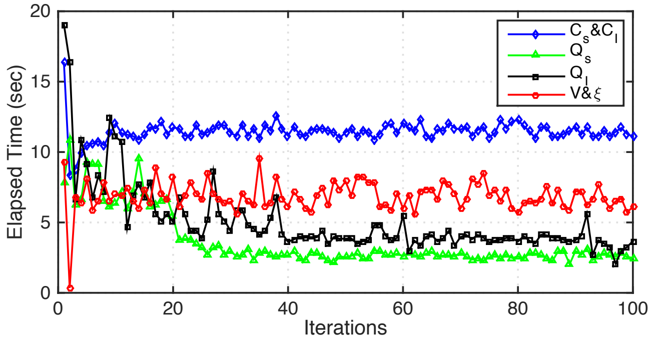

Figure 13: The elapsed time of updating four different components of each iteration in an ACDL training.

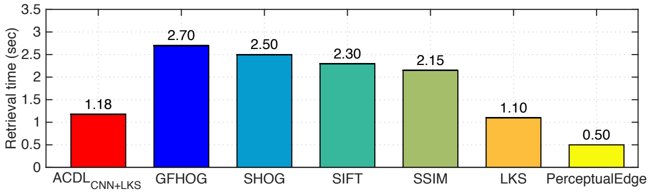

Figure 14: Comparison of average retrieval time.

we consider ACDL$_{\text{CNN+LKS}}$) contains three steps for the retrieval: (i) feature extraction from a query sketch sample using LKS, (ii) dictionary mapping to obtain a new feature representation and (iii) query the image features database with k-NN. The last two steps are very fast, and the feature extraction using LKS takes around 1 second. The average retrieval time for each query sample is around 1.18 seconds. PerceptualEdge method achieves the best retrieval speed, as it uses only two steps namely the HOG feature extraction and direct matching. The retrieval speed of ours is comparable to the LKS method, and almost 2 times faster than GFHOG, SHOG, SIFT and SSIM, which first extract features, and then construct bag-of-words features and finally perform the retrieval. The reason is that the step of constructing the bag-of-words features is more time consuming than the dictionary mapping step.

5. CONCLUSIONS

In this paper we introduce academic coupled dictionary learning (ACDL) to tackle the SBIR task. In our approach the learning pace is determined from the reconstruction error, the cross-modal code coherence and the prior knowledge suggested by the modality-specific partial curricula. Importantly, the proposed framework naturally handles different descriptors for the sketch and the image domains. Therefore, domain-specific discriminative feature representations (*e.g.*, CNN features for images) are considered, overcoming the limitations of previous works. Extensive evaluations on three publicly available datasets show that ACDL outperforms state-of-the-art approaches for SBIR. Since the method is able to learn from a small training subset and handle partial curricula, the proposed method is applicable on very large-scale datasets. Future works include investigating cross-domain deep architectures [40] able to incorporate prior information from the partial curricula.

6. ACKNOWLEDGMENTS

This research has been partially supported by the EU projet xLiMe.

7. REFERENCES

[1] B. Alexe, T. Deselaers, and V. Ferrari. Measuring the objectness of image windows. *IEEE Trans. on PAMI*, 34(11):2189–2202, 2012.

[2] A. Beck and M. Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM J. Imag. Sciences*, 2(1):183–202, 2009.

[3] Y. Bengio, J. Louradour, R. Collobert, and J. Weston. Curriculum learning. In *ICML*, 2009.

[4] A. D. Bimbo and P. Pala. Visual image retrieval by elastic matching of user sketches. *IEEE Trans. on PAMI*, 19(2):121–132, 1997.

[5] Y. Cao, H. Wang, C. Wang, Z. Li, L. Zhang, and L. Zhang. Mindfinder: interactive sketch-based image search on millions of images. In *ACM Multimedia*, 2010.

[6] A. Chalechale, G. Naghdy, and A. Mertins. Sketch-based image matching using angular partitioning. *IEEE Trans. on Systems, Man, and Cybernetics*, 35(1):28–41, 2005.

[7] M. Eitz, J. Hays, and M. Alexa. How do humans sketch objects? *ACM Trans. on Graphics*, 31(4):44, 2012.

[8] M. Eitz, K. Hildebrand, T. Boubekeur, and M. Alexa. An evaluation of descriptors for large-scale image retrieval from sketched feature lines. *Computers & Graphics*, 34(5):482–498, 2010.

[9] M. Eitz, K. Hildebrand, T. Boubekeur, and M. Alexa. Sketch-based image retrieval: Benchmark and bag-of-features descriptors *IEEE Trans. on Visualization and Computer Graphics*, 17(11):1624–1636, 2011.

[10] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman. The pascal visual object classes (VOC) challenge. *IJCV*, 88(2):303–338, 2010.

[11] J. Guo, C. Wang, and H. Chao. Building effective representations for sketch recognition. In *AAAI*, 2015.

[12] R. Hu and J. Collomosse. A performance evaluation of gradient field hog descriptor for sketch based image retrieval. *CVIU*, 117(7):790–806, 2013.

[13] D.-A. Huang and Y.-C. F. Wang. Coupled dictionary and feature space learning with applications to cross-domain image synthesis and recognition. In *ICCV*, 2013.

[14] Y. Jia. Caffe: An open source convolutional architecture for fast feature embedding, 2013.

[15] L. Jiang, D. Meng, T. Mitamura, and A. G. Hauptmann. Easy samples first: Self-paced reranking for zero-example multimedia search. In *ACM Multimedia*, 2014.

[16] L. Jiang, D. Meng, S.-I. Yu, Z. Lan, S. Shan, and A. Hauptmann Self-paced learning with diversity. In *NIPS*, 2014.

[17] L. Jiang, D. Meng, Q. Zhao, S. Shan, and A. G. Hauptmann. Self-paced curriculum learning. In *AAAI*, 2015.

[18] M. P. Kumar, B. Packer, and D. Koller. Self-paced learning for latent variable models. In *NIPS*, 2010.

[19] H. Lee, A. Battle, R. Raina, and A. Y. Ng. Efficient sparse coding algorithms. In *NIPS*, 2006.

[20] Y. J. Lee and K. Grauman. Learning the easy things first: Self-paced visual category discovery. In *CVPR*, 2011.

[21] Y. Li, T. M. Hospedales, Y.-Z. Song, and S. Gong. Fine-grained sketch-based image retrieval by matching deformable part models. In *BMVC*, 2014.

[22] Y.-L. Lin, C.-Y. Huang, H.-J. Wang, and W.-C. Hsu. 3d sub-query expansion for improving sketch-based multi-view image retrieval. In *ICCV*, 2013.

[23] J. Mairal, F. Bach, J. Ponce, G. Sapiro, and A. Zisserman. Non-local sparse models for image restoration. In *ICCV*, 2009.

[24] Y. Matsui. Challenge for manga processing: Sketch-based manga retrieval. In *ACM Multimedia*, 2015.

[25] A. Pentina, V. Sharmanska, and C. H. Lampert. Curriculum learning of multiple tasks. *CVPR*, 2015.

[26] Y. Qi, Y.-Z. Song, T. Xiang, H. Zhang, T. Hospedales, Y. Li, and J. Guo. Making better use of edges via perceptual grouping. In *CVPR*, 2015.

[27] J. M. Saavedra, J. M. Barrios, and S. Orand. Sketch based image retrieval using learned keyshapes (LKS). In *BMVC*, 2015.

[28] J. M. Saavedra and B. Bustos. An improved histogram of edge local orientations for sketch-based image retrieval. In *Pattern Recognition*, pages 432–441. Springer, 2010.

[29] J. M. Saavedra and B. Bustos. Sketch-based image retrieval using keyshapes. *Multimedia Tools and Applications*, 73(3):2033–2062, 2014.

[30] A. Sharma and D. W. Jacobs. Bypassing synthesis: Pls for face recognition with pose, low-resolution and sketch. In *CVPR*, 2011.

[31] X. Sun, C. Wang, C. Xu, and L. Zhang. Indexing billions of images for sketch-based retrieval. In *ACM Multimedia*, 2013.

[32] J. S. Supancic and D. Ramanan. Self-paced learning for long-term tracking. In *CVPR*, 2013.

[33] X. Tang and X. Wang. Face sketch recognition. *IEEE Trans. Cir. Sys. Video Tech.*, 14(1):50–57, 2004.

[34] Y. Tang, Y.-B. Yang, and Y. Gao. Self-paced dictionary learning for image classification. In *ACM Multimedia*, 2012.

[35] J. B. Tenenbaum and W. T. Freeman. Separating style and content with bilinear models. *Neural Computation*, 12(6):1247–1283, 2000.

[36] S. Wang, L. Zhang, Y. Liang, and Q. Pan. Semi-coupled dictionary learning with applications to image super-resolution and photo-sketch synthesis. In *CVPR*, 2012.

[37] X. Wang and X. Tang. Face photo-sketch synthesis and recognition. *IEEE Trans. on PAMI*, 31(11):1955–1967, 2009.

[38] C. Xiao, C. Wang, L. Zhang, and L. Zhang. Sketch-based image retrieval via shape words. In *ACM ICMR*, 2015.

[39] C. Xu, D. Tao, and C. Xu. Multi-view self-paced learning for clustering. In *IJCAI*, 2015.

[40] D. Xu, E. Ricci, Y. Yan, J. Song, and N. Sebe. Learning deep representations of appearance and motion for anomalous event detection. In *BMVC*, 2015.

[41] Y. Yan, Y. Yang, H. Shen, D. Meng, G. Liu, A. Hauptmann, and N. Sebe. Complex event detection via event oriented dictionary learning. In *AAAI*, 2015.

[42] J. Yang, J. Wright, T. S. Huang, and Y. Ma. Image super-resolution via sparse representation. *IEEE Trans. on Image Processing*, 19(11):2861–2873, 2010.

[43] S. M. Yoon, G.-J. Yoon, and T. Schreck. User-drawn sketch-based 3d object retrievalusing sparse coding. *Multimedia Tools App.*, 74(13):4707–4722, 2015.

[44] X. Zhai, Y. Peng, and J. Xiao. Heterogeneous metric learning with joint graph regularization for cross-media retrieval. In *AAAI*, 2013.

[45] Q. Zhao, D. Meng, L. Jiang, Q. Xie, Z. Xu, and A. G. Hauptmann. Self-paced learning for matrix factorization. In *AAAI*, 2015.