The CNN News Footage Datasets: Enabling Supervision in Image Retrieval

Çağdaş Bilen, Joaquin Zepeda and Patrick Pérez
Technicolor
975 avenue des Champs Blancs, CS 17616, 35576 Cesson Sévigné, France
{cagdas.bilen, joaquin.zepeda, patrick.perez}@technicolor.com

***Abstract*—Image retrieval in large image databases is an important problem that drives a number of applications. Yet the use of supervised approaches that address this problem has been limited due to the lack of large labeled datasets for training. Hence, in this paper we introduce two new datasets composed of images extracted from publicly available videos from the Cable News Network (CNN). The proposed datasets are particularly suited to supervised learning for image retrieval and are larger than any other existing dataset of a similar nature. The datasets are further provided with a set of pre-computed, state-of-the-art image feature vectors, as well as baseline results. In order to facilitate research in this important topic, we also detail a generic, supervised learning formulation for image retrieval and a related stochastic solver.**

I. Introduction

The dawn of the handheld device has brought along with it an explosion of acquisition and storage of image and video content, and accordingly a need to automatically search and retrieve this content based solely on the available low-level pixel information. Amongst the various existing *image search* methods that address this need, one can distinguish roughly between two main tasks: *i) image classification*, which consists of determining the presence or absence of a visual class (*e.g.*, *cat*) in a new and never-seen image given a set of examples of that class, and *ii) image retrieval*, which consists of retrieving images of a specific scene (or object) given an example image, despite possibly large photometric, perspective or background differences.

Particularly over the last two decades, learning algorithms have became crucial enablers of the image search tasks described above. A paramount example of this is the method of Sivic and Zisserman [1], which used a K-means-learned codebook to derive the Bag-of-Words (BoW) image feature. BoW soon became a crucial component of various image retrieval and classification systems, and to this day, the related K-means-derived inverted file index is an integral part of high-speed image retrieval systems [2]. The VLAD feature vector [3], an extension of BoW likewise based on K-means codebooks, further leveraged spatially-dependent rotations learned via PCA. VLAD and the related Fisher feature vector [4] achieve state-of-the-art performance in image retrieval, with the latter substituting the codebook with a learned Gaussian Mixture Model (GMM). Note that all of these methods (K-means, PCA, and GMM) estimate models of the distribution of the training images (or rather, of low-level feature vectors extracted from these), and hence do not require any sort of human annotation of the training images. Such learning methods are termed *unsupervised*.

In *supervised learning*, on the other hand, the training set is embedded with some notion of human understanding that the system being trained needs to reflect. A paramount example of this is image classification, where each training example is an $(\mathbf{x}, y)$ pair, with $\mathbf{x} \in \mathbb{R}^d$ a low-level feature (*e.g.*, a color-histogram) extracted automatically from the image, and $y \in \{-1, 1\}$ a human-generated annotation indicating the presence or absence of a class. An impressive recent example of the merits of supervised learning in classification is the resurgence, starting in 2012 [5], of deeply stacked neural networks.

Yet the image retrieval task continues to be an exception to the exploitation of supervised learning. One of the main reasons for this is the lack of adequate training datasets: While very large datasets consisting of up to millions of images and thousands of classes indeed exist for face verification [6], [7] and image classification [8], the existing datasets for image retrieval are small and meant only to be used as a comparison tool.

The Oxford buildings dataset [9], for example, contains only $5,062$ images of 11 different scenes. The INRIA Holidays dataset [10] contains only $1,092$ images of about 500 scenes. An earlier dataset, the UKBench dataset [11], is larger and more diverse, but its images contain a single, well-centered object and hence most recent methods already perform perfectly on this dataset.

The aim of the present work is thus to introduce two new datasets that will enable the use of supervised learning for image retrieval. The first of those, the CNN News Footage Dataset, consists of keyframes derived from a large amount of public access videos from the Cable News Network (CNN). News footage was chosen due to its wide variety of content. The second dataset, Extended CNN News Footage Dataset, includes more images which are artificially generated from the images of CNN News Footage Dataset by randomly rotating, translating, cropping and scaling to make the task of ranking more challenging. Our datasets further include a full set of VLAD feature vectors of various dimensionalities.[1]

[1]Both the CNN News Footage and Extended CNN News Footage databases can be obtained from the website: http://www.technicolor.com/en/patrick-perez

Besides the main contribution discussed above, we make two supplementary contributions. First, we discuss a generic framework to carry out learning for image retrieval using Stochastic Gradient Descent (SGD) that can be exploited by specific research endeavours relying on our dataset. And secondly, we provide baseline results on our dataset with an unsupervised and supervised metric learning algorithm from the literature.

II. Learning For Image Retrieval

In this section we present a generic learning framework that can be used to carry out supervised learning for image retrieval using the introduced CNN News Footage dataset. We first present a learning problem that is well suited for the image retrieval task and subsequently discuss a generic Stochastic Gradient Descent (SGD) algorithm to solve this problem.

A. Learning objectives

Similarly to other image retrieval datasets, the CNN News Footage datasets we introduce are organized into groups of matching images. From these groups, we derive pairs of matching or non-matching images. Such pairs can be represented using $\mathcal{P} = (i, j, y)$ tuples, where i and j are image indices, and the *annotation* $y \in \{1, -1\}$ denotes whether the indicated pair of images $\mathbf{I}_i$ and $\mathbf{I}_j$ match ($y = 1$) or do not match ($y = -1$). We refer to pairs having $y = 1$ as *positive* pairs, and those having $y = -1$ as *negative* pairs.

Using this organization of the training set, one is interested in learning methods that enforce that distances between images of matching pairs are lower than distances between images of non-matching pairs. We let $d_{\mathbf{\Theta}}(\mathbf{I}_i, \mathbf{I}_j)$ denote a distance function used to compare images and which depends on a set of parametres $\mathbf{\Theta}$. This representation can correpond to a wide range of situations. For example, $\mathbf{\Theta}$ can represent a positive semi-definite (p.s.d.) Mahalanobis matrix, in which case, letting $f(\mathbf{I}) \in \mathbb{R}^d$ denote a vector representation of an image $\mathbf{I}$, the distance function $d_{\mathbf{\Theta}}$ takes the following form:

$$d_{\mathbf{\Theta}}(\mathbf{I}_i, \mathbf{I}_j) = \left(f(\mathbf{I}_i) - f(\mathbf{I}_j)\right)^\top \mathbf{\Theta} \left(f(\mathbf{I}_i) - f(\mathbf{I}_j)\right), \quad \text{for p.s.d. } \mathbf{\Theta}. \tag{1}$$

Alternatively, the image representation function f can itself depend on a set of parameters $\mathbf{\Theta}$, and the metric computation can take a simpler form, *e.g.*,

$$d_{\mathbf{\Theta}}(\mathbf{I}_i, \mathbf{I}_j) = \|f(\mathbf{I}_i; \mathbf{\Theta}) - f(\mathbf{I}_j; \mathbf{\Theta})\|_2. \tag{2}$$

Possible formulations for the representation function f include the GMM or VLAD representations, where $\mathbf{\Theta}$ would correspond to the codebook or GMM parameters, or deep convolutional architectures where $\mathbf{\Theta}$ would represent the convolutional filters.

a) Representation and metric decomposition: More generally, it is possible to have approaches that combine metric models such as (1) and feature representation models such as (2). It is convenient in such cases to express $d_{\mathbf{\Theta}}$ in terms of *i)* an *image representation function*

$$f : \mathcal{I} \times \mathcal{T}_1 \to \mathbb{R}^d \tag{3}$$

that maps the images $\mathbf{I} \in \mathcal{I}$ and model parameters $\mathbf{\Theta}_1 \in \mathcal{T}_1$ to a d-dimensional vector representation of the image, and *ii)* a metric

$$g : \mathbb{R}^d \times \mathbb{R}^d \times \mathcal{T}_2 \to \mathbb{R}^+ \tag{4}$$

that assigns a positive scalar to two input image feature vectors given the model parameters $\mathbf{\Theta}_2 \in \mathcal{T}_2$. Accordingly, $d_{\mathbf{\Theta}}$ can be written as follows, where $\mathbf{\Theta} = (\mathbf{\Theta}_1, \mathbf{\Theta}_2)$:

$$d_{\mathbf{\Theta}}(\mathbf{I}_i, \mathbf{I}_j) = g\left(f(\mathbf{I}_i; \mathbf{\Theta}_1), f(\mathbf{I}_j; \mathbf{\Theta}_1); \mathbf{\Theta}_2\right). \tag{5}$$

This formulation is compatible with a wide range of existing metric learning methods and image feature extraction methods, whether supervised or unsupervised.

b) Learning $\mathbf{\Theta}$: In order to favor the generalization power of learned models, it is desirable that distances for matching pairs are lower than those for non-matching pairs by some non-negligible margin $b > 0$. Using the overloaded notation

$$d_{\mathbf{\Theta}}(\mathcal{P}) \triangleq d_{\mathbf{\Theta}}(\mathbf{I}_i, \mathbf{I}_j) \tag{6}$$

for convenience, this margin-compliant distance constraint can be represented as follows, where $\mathcal{P}_m = (i_m, j_m, y_m = 1)$ contains a positive pair and $\mathcal{P}_n = (i_n, j_n, y_n = -1)$ a negative pair:

$$d_{\mathbf{\Theta}}(\mathcal{P}_n) - d_{\mathbf{\Theta}}(\mathcal{P}_m) - b \geq 0. \tag{7}$$

In practice, it might not be possible to enforce the above constraints for all possible $\mathcal{P}_m$ and $\mathcal{P}_n$. Similarly to the approach used to train Support Vector Machines (SVMs) over non-linearly-separable training sets [12], one can relax some of the constraints by means of learned, non-negative slack variables $\xi_{mn} \geq 0$ using

$$d_{\mathbf{\Theta}}(\mathcal{P}_n) - d_{\mathbf{\Theta}}(\mathcal{P}_m) - b + \xi_{mn} \geq 0. \tag{8}$$

A sensible optimization strategy hence amounts to reducing the number of active slack variables (*i.e.*, making the variables $\{\xi_{mn}\}_{m,n}$ sparse), as well as minimizing the magnitude of those that are active. Letting

$$\mathcal{M} \triangleq \{\mathcal{P}_k \text{ s.t. } y_k = 1\} \text{ and } \mathcal{N} \triangleq \{\mathcal{P}_k \text{ s.t. } y_k = -1\}, \tag{9}$$

denote, respectively, the set of all positive pairs and all negative pairs, a relaxed formulation of this strategy relying on the ℓ_1 norm of $\{\xi_{mn}\}_{m,n}$ and enjoying convexity in the ξ_{mn} is

$$\begin{aligned} \min_{\mathbf{\Theta}, \{\xi_{mn}\}} \quad & \lambda \Omega\left(\mathbf{\Theta}\right) \quad + \quad \frac{1}{N} \sum_{\substack{m \in \mathcal{M} \\ n \in \mathcal{N}}} \xi_{mn} \\ & d_{\mathbf{\Theta}}(\mathcal{P}_n) - d_{\mathbf{\Theta}}(\mathcal{P}_m) - b + \xi_{mn} \geq 0, \quad \xi_{mn} \geq 0 \end{aligned} \tag{10}$$

where N is the total number of terms in the summation and Ω is a regularization function with scalar penalty weight $\lambda \geq 0$. The two constraints on the variables ξ_{mn} can be used to express ξ_{mn} in closed form by means of the hinge loss

$$\ell_b(x) \triangleq \max(0, b - x) \tag{11}$$

as follows:

$$\xi_{mn} = \ell_b(d_{\mathbf{\Theta}}(\mathcal{P}_n) - d_{\mathbf{\Theta}}(\mathcal{P}_m)). \tag{12}$$

Accordingly, the resulting form of (10) is

$$\min_{\mathbf{\Theta}} \ \lambda\Omega\left(\mathbf{\Theta}\right) \ + \ \frac{1}{N}\sum_{\substack{m\in\mathcal{M}\\ n\in\mathcal{N}}} \ell_b\left(d_{\mathbf{\Theta}}(\mathcal{P}_n) - d_{\mathbf{\Theta}}(\mathcal{P}_m)\right). \quad (13)$$

For completeness we note that one could restrict the summation over n to the subset

$$\mathcal{N}_m = \{\mathcal{P}_k \text{ s.t. } i_k = i_m, y_k = -1\} \quad (14)$$

derived from image index i_m in $\mathcal{P}_m = (i_m, j_m, 1)$ or, alternatively, restrict the summation over m to the analogous set

$$\mathcal{M}_n = \{\mathcal{P}_k \text{ s.t. } i_k = i_n, y_k = 1\} \quad (15)$$

derived from $\mathcal{P}_n = (i_n, j_n, -1)$. Using either of the above restrictions covers the case where the groups are organized into triplets of the form (i, i_+, i_-), with i being the index of an arbitrary image, and i_+ (respectively, i_-) the index of a matching (non-matching) image [13].

B. Related methods

The formulation in (13) is a generic formulation that covers various existing metric and representation learning methods, possibly with minor variations.

One potential problem with (13) is that, particularly for feature spaces $\mathbb{R}^d$ of high dimension d, the values of distances $d_\Theta(\mathcal{P})$ will vary greatly in magnitude across space, and this will artificially enhance the contribution of certain pairs to the objective function. This problem can be addressed by means of spatially-dependent normalization multipliers $w(\mathcal{P}_m, \mathcal{P}_n)$ that weigh the ℓ_b term inside (13) and that can be subsumed within g in our formulation.

The approach of [14] considers instead weights $w(\mathcal{P}_m, \mathcal{N}_m)$ that depend on all possible negative pairs $\mathcal{N}_m$ derived from image index i_m in $\mathcal{P}_m = (i_m, j_m, y_m)$:

$$w(\mathcal{P}_m, \mathcal{N}_m) = \frac{\sum_{k=1}^{r_m} \frac{1}{k}}{r_m}, \quad (16)$$

where $r_m = |\{\mathcal{P}_n \in \mathcal{N}_m \text{ s.t. } d_{\mathbf{\Theta}}(\mathcal{P}_n) < d_{\mathbf{\Theta}}(\mathcal{P}_m)\}|$ is the number of pairs from $\mathcal{N}_m$ with distance lower than that for $\mathcal{P}_m$.

The method of [15] uses a small variation of (13) to learn a codebook for the VLAD image representation. The main difference in their objective function is that the summation over $m \in \mathcal{M}$ in (13) is substituted by a minimization over $m \in \mathcal{M}_n$. This adaptation is intented to mitigate the noisiness of the annotations y that is a consequence of their automated training set compilation strategy. The end result is that the objective in (13) will depend only on those positive pairs $\mathcal{P}_m = (i_m, j_m, 1)$ such that the image (with index) j_m is closest in feature space to image i_m than all other images j, and hence most likely to be a correct match.

C. Stochastic Gradient Descent (SGD)

Stochastic Gradient Descent (SGD) is an optimization method that can be applied to empirical risks of the form

$$\frac{1}{N}\sum_{i=1}^{N} \phi\left(\mathbf{\Theta}, \mathcal{S}_i\right), \quad (17)$$

where $\mathcal{S}_i$ is a sample from the training set and $\mathbf{\Theta}$ is set of model parameters being learned. At iteration t, SGD proceeds by drawing a random example $\mathcal{S}_{i_t}$ from the training set and updating the current estimate $\mathbf{\Theta}_t$ of the parameters using [2]

$$\mathbf{\Theta}_{t+1} = \mathbf{\Theta}_t - \gamma_t \nabla_{\mathbf{\Theta}} \phi\left(\mathbf{\Theta}, \mathcal{S}_{i_t}\right), \quad (18)$$

where $\gamma_t = \gamma/(t + t_0)$ is the *learning rate* [16], with coefficients γ and t_0 set by means of cross-validation.

For completeness, we give generic gradient expressions enabling the application of (18) to the problem in (13) for distances of the form (5), where the training examples $\mathcal{S}_i$ can be taken to be $(\mathcal{P}_m, \mathcal{P}_n)$-pairs and

$$\phi\left(\mathbf{\Theta}, (\mathcal{P}_m, \mathcal{P}_n)\right) = \lambda\Omega\left(\mathbf{\Theta}\right) + \ell_b\left(d_{\mathbf{\Theta}}(\mathcal{P}_n) - d_{\mathbf{\Theta}}(\mathcal{P}_m)\right). \quad (19)$$

Accordingly, letting $[\![\cdot]\!]$ evaluate to 1 if the condition is true and 0 otherwise, the (sub-)gradient required in (18) is

$$\nabla_{\mathbf{\Theta}}\phi = \lambda\nabla_{\mathbf{\Theta}}\Omega - [\![b - z \geq 0]\!]\big(\nabla_{\mathbf{\Theta}} d_{\mathbf{\Theta}}(\mathcal{P}_n) - \nabla_{\mathbf{\Theta}} d_{\mathbf{\Theta}}(\mathcal{P}_m)\big), \quad (20)$$

where $z = d_{\mathbf{\Theta}}(\mathcal{P}_n) - d_{\mathbf{\Theta}}(\mathcal{P}_m)$ and the gradients of $d_{\mathbf{\Theta}}$ can be assembled from the partial gradients $\nabla_{\mathbf{\Theta}_1} d_{\mathbf{\Theta}}(\mathcal{P}) = \nabla_{\mathbf{\Theta}_1} g$ and, letting $f_k \triangleq f(\mathbf{I}_k; \mathbf{\Theta}_2)$ and $\mathcal{P} = (i, j, y)$,

$$\nabla_{\mathbf{\Theta}_2} d_{\mathbf{\Theta}}(\mathcal{P}) = \left.\frac{\partial g(x, f_j; \mathbf{\Theta}_2)}{\partial x}\right|_{f_i} \cdot \frac{\partial f_i}{\partial \mathbf{\Theta}_2} + \left.\frac{\partial g(f_i, x; \mathbf{\Theta}_2)}{\partial x}\right|_{f_j} \cdot \frac{\partial f_j}{\partial \mathbf{\Theta}_2}. \quad (21)$$

III. The CNN News Footage Datasets

In this section, we provide the details on how the images of the CNN News Footage Dataset are selected and processed and how the Extended CNN News Footage Dataset is generated.

A. Image Source and Characteristics

The images in the database are taken from the videos of Cable News Network (CNN) news reports that are publicly available from Internet Archive[3] and are used in online demo of EU Project AXES.[4] These videos are chosen particularly because of their variety in content in addition to being publicly available for use. The keyframes from the videos of CNN news reports of a total of 522 days from the year 2007 to 2011 are extracted and processed as described in Section III-B. The original keyframes have the resolution of 320×240 in the videos from earlier years with analog broadcast and 400×224 in the videos from later years with digital broadcast. The video

[2] We use the convention that gradients are row vectors and, for $f : \mathbb{R}^a \to \mathbb{R}^b, \mathbf{x} \mapsto f(\mathbf{x})$, $\frac{\partial f}{\partial \mathbf{x}} \in \mathbb{R}^{b\times a}$ is the Jacobian matrix. For consistency, $\mathbf{\Theta}$ can be thought of as a row vector.

[3] https://archive.org/details/TV-CNN

[4] http://www.axes-project.eu/?page_id=2310

Fig. 1. **Positive pairs**. Some examples of matching pairs formed by images from the groups of CNN News Footage dataset. They span a variety of scenes, from close-ups to full shots, from empty to cluttered background, and of intra-pair variations, from small camera/object movements to large scene changes due to pose variations or object movements.

Fig. 2. **Geometric processing of images**. Examples of how the images are processed in the dataset preparation: The images are cropped (as indicated with blue box) to exclude text overlays; For temporally close images, the common field of view (indicated with red boxes) is tentatively obtained through homographic matching and similarity is decided for this pair based on the confidence of the homography and the relative size of the shared region.

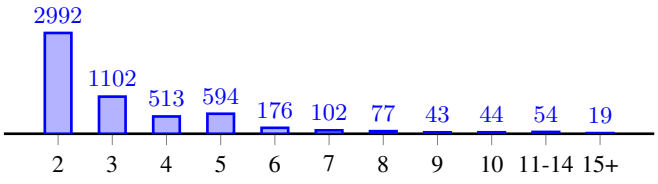

Fig. 3. **Image groups statistics**. Size histogram for the 5,417 image groups that form the CNN News Footage dataset.

content include studio discussions, street interviews, and the large variety of footage that support news reports.

B. Annotation Methodology

The extracted keyframes of the news footage are processed and annotated by following the steps:

1) The images are sorted chronologically and local features are extracted from each image (excluding the ones in bottom regions where text overlays appear). The features of each image are matched to the features of other images that are chronologically close, in order to find the homography and the common areas between the image pairs as shown in Figure 2. The pairs that have a valid transformation and large enough common area are selected to be in the same similarity group. Identical images are also removed.
2) All the images are cropped to a size of 320×180 as shown in Figure 2 so as to remove the text overlays that appear in the bottom of the picture in the news reports.
3) All the images are manually filtered to exclude the images with:
   - Synthetic imagery (such as weather reports or split screen interviews);
   - Blurry or low quality images;
   - Images that appear too frequently (such as the news anchors recorded in the studio).
4) The image groups are manually checked to join the groups with very similar images and remove irrelevant images from the groups. Finally the groups with a single image are also removed.

The finalized database is composed of 17038 images organized within 5417 groups. The histogram of the number of images in these groups is shown in Figure 3. All images within a group are considered to be mutually similar (or matching). A set of positive (and negative) image pairs (or a set of image triplets) can be generated from these groups for the purpose of training and testing supervised retrieval algorithms. Some example pairs from a number of groups are shown in Figure 1.

C. The Extended Database

In order to increase the number of images and to make the image retrieval task more challenging, we have also created an extended version of the CNN News Footage Dataset. The Extended CNN News Footage Dataset is created by generating new images from the existing images in CNN News Footage so as to have 10 or more images in each image group (including the original images of CNN News Footage database). The additional images are generated by randomly (i) rotating, (ii) translating, (iii) cropping and (iv) scaling the original images. The transformation effects are applied jointly so that the image is randomly rotated, translated and cropped with the constraint that the final image is always within the original image (no black borders in the randomly generated images). Finally it is scaled up to so that larger side is 320 pixels. The Extended CNN News Footage database is composed of 66728 images organized within 5417 groups and it is significantly

	CNN				Extended CNN			
d	16	32	64	128	16	32	64	128
PCA	0.80	0.89	0.94	0.97	0.04	0.04	0.06	0.06
ML	0.84	0.93	0.95	0.98	0.10	0.13	0.15	0.17

TABLE I
BASELINE MAP RESULTS FOR THE CNN NEWS FOOTAGE AND EXTENDED CNN NEWS FOOTAGE DATASETS USING VARIOUS DIMENSIONS d OF PROJECTION MATRICES.

	CNN				Extended CNN			
d	16	32	64	128	16	32	64	128
PCA	0.83	0.92	0.95	0.97	0.10	0.13	0.21	0.25
ML	0.87	0.94	0.96	0.98	0.33	0.42	0.49	0.53

TABLE II
BASELINE P_1 RESULTS FOR THE CNN NEWS FOOTAGE AND EXTENDED CNN NEWS FOOTAGE DATASETS USING VARIOUS DIMENSIONS d OF PROJECTION MATRICES.

more challenging than CNN News Footage database for image retrieval task.

IV. BASELINE RESULTS

In this section we present baseline results for our proposed CNN News Footage and Extended CNN News Footage datasets. These results are meant as a comparison basis for future image retrieval methods relying on supervised learning that exploit our dataset. For this purpose 500 randomly selected groups with a total of 1635 images are separated as a test set, and the images in the remaining groups are used to randomly generate positive and negative pairs for training.

We compute results using the Mahalanobis metric learning (ML) in (1) under the learning objective in (13) but with weights as specified in (16) [14]. For the image representation function f, we use the VLAD representation [3] based on local SIFT descriptors extracted densely over a regular grid at three different scales. We restrict the rank of p.s.d. matrix $\mathbf{\Theta}$ to a fixed value d. This amounts to computing a d-dimensional projection of the VLAD representation, and we provide baseline results for various d.

As a performance measure, we use mean Average Precision (mAP), which we now describe: For a given test query image $\mathbf{I}$, the other test images are ranked according to learned distance $d_{\mathbf{\Theta}}(\mathbf{I}, \cdot)$. We let $T_k \leq k$ denote the number of true matching images within the k top-ranked images. We also let Q denote the total number of true matches for that query image. Accordingly, precision P_k and recall R_k at rank k are given by

$$P_k = \frac{T_k}{k}, \qquad R_k = \frac{T_k}{Q}. \tag{22}$$

The Average Precision (AP) for that query image is then the area under the curve plotting R_k vs. P_k for $k = 1, \ldots, K$, where K is the total number of images in the database. Accordingly, mAP is obtained by averaging the AP of all the query images in the test set.

In Table I and Table II, we present baseline results for various dimensions d using mAP and P_1 as performance measures. The mAP and P_1 are computed for retrieval from the entire database with query images from the test set (one query image per each group in test set). To illustrate the merits of supervised learning over unsupervised learning, we likewise present results using PCA-based projections for the same dimensions.

V. CONCLUSION

In this work, we introduce two new datasets tailored for the image retrieval task wherein images match if they contain the same scene or object, albeit under potentially wide variations in pose. Up to now, datasets for image retrieval have been too small and used only as a comparison tool. The datasets we present, on the other hand, is large enough to enable supervised learning, a rich vein until now untapped by methods addressing the image retrieval task. Along with our datasets, we provide a set of image feature vectors to enable quick prototyping by future research efforts, as well as baseline results for comparison purposes. We further present a generic supervised learning method for the retrieval task including the generic problem formulation and related stochastic solver.

REFERENCES

[1] J. Sivic and A. Zisserman, "Video Google: A text retrieval approach to object matching in videos," in *International Conference on Computer Vision*, 2003. [Online]. Available: http://ieeexplore.ieee.org/xpls/abs{_}all.jsp?arnumber=1238663

[2] Y. Kalantidis and Y. Avrithis, "Locally Optimized Product Quantization for Approximate Nearest Neighbor Search," *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, vol. 1, 2014. [Online]. Available: http://image.ntua.gr/iva/files/lopq.pdf

[3] J. Delhumeau, P.-H. Gosselin, H. Jégou, and P. Pérez, "Revisiting the VLAD image representation," in *Proceedings of ACM International Conference on Multimedia*, vol. 21. New York, New York, USA: ACM Press, 2013, pp. 653–656. [Online]. Available: http://dl.acm.org/citation.cfm?doid=2502081.2502171

[4] F. Perronnin, Y. Liu, J. Sánchez, and H. Poirier, "Large-scale Image Retrieval with Compressed Fisher Vectors," in *Computer Vision and Pattern Recognition*. Ieee, jun 2010, pp. 3384–3391. [Online]. Available: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=5540009

[5] A. Krizhevsky, I. Sutskever, and G. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks," in *Neural Information Processing Systems*, 2012, pp. 1–9.

[6] M. Guillaumin, J. Verbeek, and C. Schmid, "Is that you? Metric learning approaches for face identification," in *International Conference on Computer Vision*, 2009. [Online]. Available: http://hal.inria.fr/inria-00439290/en

[7] M. Everingham, J. Sivic, and A. Zisserman, "Hello! My name is... Buffy Automatic naming of characters in TV video," 2006. [Online]. Available: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.104.1500

[8] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, C. V. Jan, J. Krause, and S. Ma, "ImageNet Large Scale Visual Recognition Challenge," in *arXiv*, 2014. [Online]. Available: http://image-net.org/challenges/LSVRC/2015/

[9] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman, "Object retrieval with large vocabularies and fast spatial matching," in *Computer Vision and Pattern Recognition*, 2007.

[10] Hervé Jégou, M. Douze, and C. Schmid, "Hamming Embedding and Weak Geometry Consistency for Large Scale Image Search," in *European Conference on Computer Vision*, 2008. [Online]. Available: https://lear.inrialpes.fr/{~}jegou/data.php{\#}holidays

[11] D. Nistér and H. Stewénius, "Scalable Recognition with a Vocabulary Tree," in *Computer Vision and Pattern Recognition*, 2006, pp. 2161–2168.

[12] A. Zisserman, "Lecture 2 : The SVM classifier," 2011. [Online]. Available: http://www.robots.ox.ac.uk/{~}az/lectures/ml/2011/lect2.pdf
[13] G. Chechik, V. Sharma, U. Shalit, and S. Bengio, "Large Scale Online Learning of Image Similarity Through Ranking," *Journal of Machine Learning Research*, pp. 1109–1135, 2010.
[14] D. Lim and G. Lanckriet, "Efficient Learning of Mahalanobis Metrics for Ranking," in *International Conference on Machine Learning*, vol. 32, 2014, pp. 1980–1988. [Online]. Available: http://jmlr.org/proceedings/papers/v32/lim14.html
[15] R. Arandjelović, P. Gronat, A. Torii, T. Pajdla, and J. Sivic, "NetVLAD: CNN architecture for weakly supervised place recognition," 2015. [Online]. Available: http://arxiv.org/abs/1511.07247
[16] L. Bottou, "Stochastic gradient descent tricks," in *Neural Networks: Tricks of the Trade*, 2nd ed., G. Montavon, G. Orr, and K.-R. Müller, Eds. Springer, 2012, vol. 1. [Online]. Available: http://link.springer.com/chapter/10.1007/978-3-642-35289-8{_}25