The Stanford Mobile Visual Search Data Set

Vijay Chandrasekhar[1]*, David M. Chen[1], Sam S. Tsai[1], Ngai-Man Cheung[1],
Huizhong Chen[1], Gabriel Takacs[1], Yuriy Reznik[2], Ramakrishna Vedantham[3],
Radek Grzeszczuk[3], Jeff Bach[4], Bernd Girod[1],

[1]Information Systems Laboratory, Stanford University, Stanford, CA 94305
[2]Qualcomm Inc., San Diego, CA 92121
[3]Nokia Research Center, Palo Alto, CA 94304
[4]Navteq, Chicago, IL 60606

ABSTRACT

We survey popular data sets used in computer vision literature and point out their limitations for mobile visual search applications. To overcome many of the limitations, we propose the Stanford Mobile Visual Search data set. The data set contains camera-phone images of products, CDs, books, outdoor landmarks, business cards, text documents, museum paintings and video clips. The data set has several key characteristics lacking in existing data sets: rigid objects, widely varying lighting conditions, perspective distortion, foreground and background clutter, realistic ground-truth reference data, and query data collected from heterogeneous low and high-end camera phones. We hope that the data set will help push research forward in the field of mobile visual search.

1. INTRODUCTION

Mobile phones have evolved into powerful image and video processing devices, equipped with high-resolution cameras, color displays, and hardware-accelerated graphics. They are also equipped with GPS, and connected to broadband wireless networks. All this enables a new class of applications which use the camera phone to initiate search queries about objects in visual proximity to the user (Fig 1). Such applications can be used, e.g., for identifying products, comparison shopping, finding information about movies, CDs, buildings, shops, real estate, print media or artworks. First commercial deployments of such systems include Google Goggles, Google Shopper [11], Nokia Point and Find [22], Kooaba [15], Layar [16], Ricoh iCandy [7] and Amazon Snaptell [1].

Figure 1: A snapshot of an outdoor visual search application. The system augments the viewfinder with information about the objects it recognizes in the camera phone image.

Mobile visual search applications pose a number of unique challenges. First, the system latency has to be low to support interactive queries, despite stringent bandwidth and computational constraints. One way to reduce system latency significantly is to carry out feature extraction on the mobile device, and transmit compressed feature data across the network [10]. State-of-the-art retrieval systems [14, 23] typically extract 2000-3000 affine-covariant features (Maximally Stable Extremal Regions (MSER), Hessian Affine points) from the query image. This might take several seconds on the mobile device. For feature extraction on the device to be effective, we need fast and robust interest point detection algorithms and compact descriptors. There is growing industry interest in this area, with MPEG recently launching a standardization effort [19]. It is envisioned that the standard will specify bitstream syntax of descriptors, and parts of the descriptor-extraction process needed to ensure interoperability.

Next, camera phone images tend to be of lower quality compared to digital camera images. Images that are degraded by motion blur or poor focus pose difficulties for visual recognition. However, image quality is rapidly improving with higher resolution, better optics and built-in flashes on camera phones.

Outdoor applications pose additional challenges. Current retrieval systems work best for highly textured rigid planar objects taken under controlled lighting conditions. Landmarks, on the other hand, tend to have fewer features, exhibit repetitive structures and their 3-D geometric distortions are not captured by simple affine or projective transformations. Ground truth data collection is more difficult, too. There are different ways of bootstrapping databases

*Contact Vijay Chandrasekhar at vijayc@stanford.edu.

for outdoor applications. One approach is to mine data from online collections like Flickr. However, these images tend to be poorly labelled, and include a lot of clutter. Another approach is to harness data collected by companies like Navteq, Google (StreetView) or Earthmine. In this case, the data are acquired by vehicle-mounted powerful cameras with wide-angle lenses to capture spherical panoramic images. In both cases, visual recognition is challenging because the camera phone query images are usually taken under very different lighting conditions compared to reference database images. Buildings and their surroundings (e.g., trees) tend to look different in different seasons. Shadows, pedestrians and foreground clutter are some of the other challenges in this application domain.

OCR on mobile phones enables another dimension of applications, from text input to text-based queries to a database. OCR engines work well on high quality scanned images. However, the performance of mobile OCR drops rapidly for images that are out of focus and blurry, have perspective distortion or non-ideal lighting conditions.

To improve performance of mobile visual search applications, we need good data sets that capture the most common problems that we encounter in this domain. A good data set for visual search applications should have the following characteristics:

- Should have good ground truth reference images
- Should have query images with a wide range of camera phones (flash/no-flash, auto-focus/no auto-focus)
- Should be collected under widely varying lighting conditions
- Should capture typical perspective distortions, motion blur, foreground and background clutter common to mobile visual search applications.
- Should represent different categories (e.g., buildings, books, CDs, DVDs, text documents, products)
- Should contain rigid objects so that a transformation can be estimated between the query and reference database image.

We surveyed popular data sets in the computer vision literature, and observed that they were all limited in different ways. To overcome many of the limitations in existing data sets, we propose the Stanford Mobile Visual Search (SMVS) data set that we hope will help move research forward in this field. In Section 2, we survey popular computer vision data sets, and point out their limitations. In Section 3, we propose the SMVS data set for different mobile visual search applications.

2. SURVEY OF DATA SETS

Popular computer vision data sets for evaluating image retrieval algorithms consist of a set of query images and their ground truth reference images. The number of query images typically range from a few hundred to a few thousand. The scalability of the retrieval methods is tested by retrieving the query images in the presence of “distractor” images, or images that do not belong to the data set [14, 23]. The “distractor” images are typically obtained by mining Flickr or other photo sharing websites. Here, we survey popular data sets in computer vision literature and discuss their limitations for our application. See Fig. 2 for examples from each data set, and Tab. 1 for a summary of the different data sets.

ZuBuD.

The Zurich Building (ZuBuD) dataset [12] consists of 201 buildings in Zurich, with 5 views of each building. There are 115 query images which are not contained in the database. Query and database images differ in viewpoint, but variations in illumination are rare because the different images for the same building are taken at the same time of day. The ZuBuD is considered an easy data set, with close to 100% accuracy being reported in several papers [13, 25]. Simple approaches like color histograms and descriptors based on DCT [25] yield high performance for this dataset.

Oxford Buildings.

The Oxford Buildings Datset [23] consists of 5062 images collected from Flickr by searching for particular Oxford landmarks. The collection has been manually annotated to generate a comprehensive ground truth for 11 different landmarks, each represented by 5 possible queries. This gives only a small set of 55 queries. Another problem with this data set is that completely different views of the same building are labelled by the same name. Ideally, different facades of each building should be distinguished from each other, when evaluating retrieval performance.

INRIA Holidays.

The INRIA Holidays dataset [14] is a set of images which contains personal holiday photos of the authors in [14]. The dataset includes a large variety of outdoor scene types (natural, man-made, water and fire effects). The dataset contains 500 image groups, each of which represents a distinct scene or object. The data set contains perspective distortions and clutter. However, variations in lighting are rare as the pictures are taken at the same time from each location. Also, the data set contains scenes of many non-rigid objects (fire, beaches, etc), which will not produce repeatable features, if images are taken at different times.

University of Kentucky.

The University of Kentucky (UKY) [21] consists of 2550 groups of 4 images each of objects like CD-covers, lamps, keyboards and computer equipment. Similar to ZuBuD and INRIA data sets, this data set also offers little variation in lighting conditions. Further, there is no foreground or background clutter with only the object of interest present in each image.

Image Net.

The ImageNet dataset [6] consists of images organized by nouns in the WordNet hierarchy [8]. Each node of the hierarchy is depicted by hundreds and thousands of images. E.g., Fig. 2 illustrates some images for the word “tiger”. Such a data set is useful for testing classification algorithms, but not so much for testing retrieval algorithms.

We summarize the limitations of the different data sets in Tab. 1. To overcome the limitations in these data sets, we propose the Stanford Mobile Visual Search (SMVS) data set.

3. STANFORD MOBILE VISUAL SEARCH DATA SET

We present the SMVS (version 0.9) data set in the hope that it will be useful for a wide range of visual search applications like product recognition, landmark recognition, out-

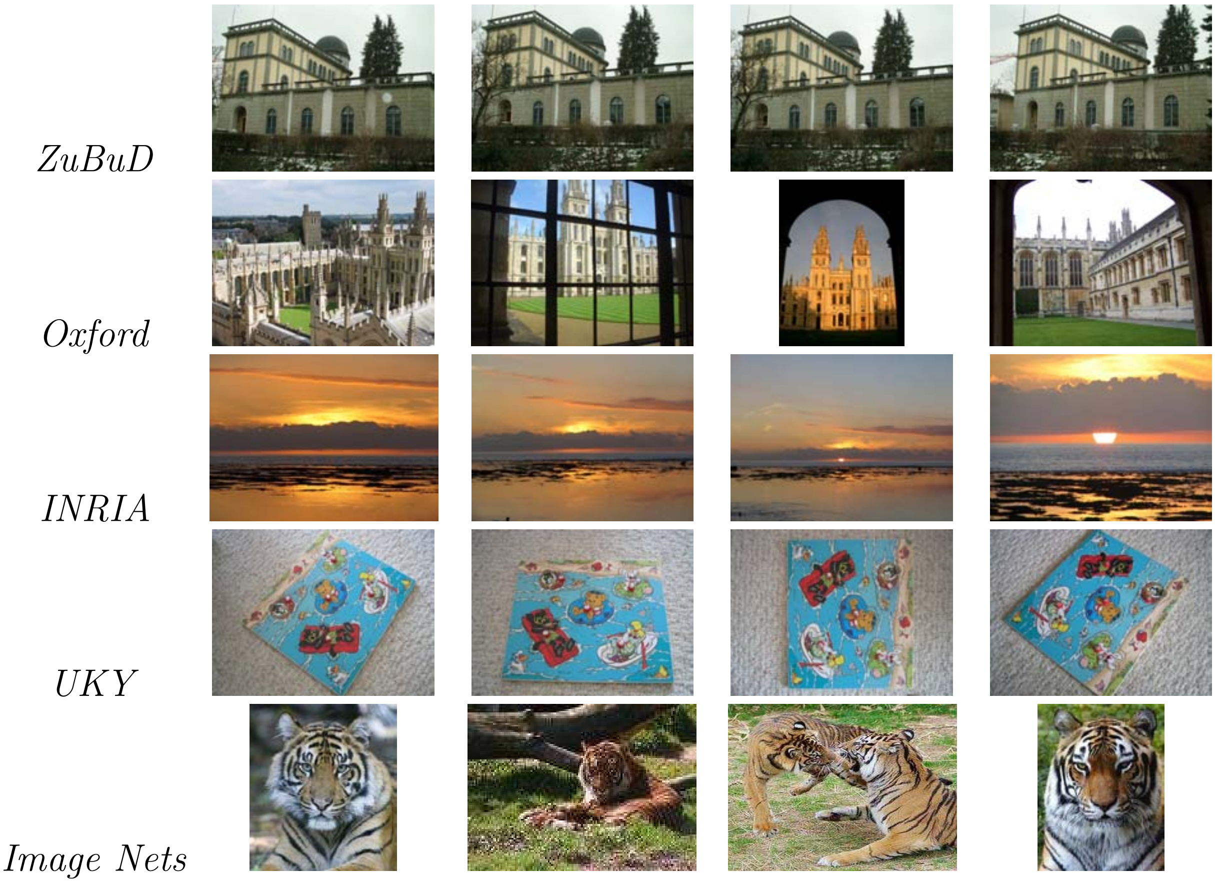


Figure 2: Limitations with popular data sets in computer vision. The left most image in each row is the database image, and the other 3 images are query images. ZuBuD, INRIA and UKY consist of images taken at the same time and location. ImageNets is not suitable for image retrieval applications. The Oxford dataset has different faades of the same building labelled with the same name.

door augmented reality [27], business card recognition, text recognition, video recognition and TV-on-the-go [5]. We collect data for several different categories: CDs, DVDs, books, software products, landmarks, business cards, text documents, museum paintings and video clips. Sample query and database images are shown in Figure 4. Current and subsequent versions of the dataset will be available at [3].

The number of database and query images for different categories is shown in Tab. 2. We provide a total 3300 query images for 1200 distinct classes across 8 image categories. Typically, a small number of query images (∼1000s) suffice to measure the performance of a retrieval system as the rest of the database can be padded with "distractor" images. Ideally, we would like to have a large distractor set for each query category. However, it is challenging to collect distractor sets for each category. Instead, we plan to release two distractor sets upon request: one containing Flickr images, and the other containing building images from Navteq. The distractor sets will be available in sets of 1K, 10K, 100K and 1M. Researchers can test scalability using these distractor data sets, or the ones provided in [23, 14]. Next, we discuss how the query and reference database images are collected, and evaluation measures that are in particular relevant for mobile applications.

Reference Database Images.

For product categories (CDs, DVDs and books), the references are clean versions of images obtained from the product websites. For landmarks, the reference images are obtained from data collected by Navteq's vehicle-mounted cameras. For video clips, the reference images are the key frame from the reference video clips. The videos contain diverse content like movie trailers, news reports, and sports. For text documents, we collect (1) reference images from [20], a website that mines the front pages of newspapers from around the world, and (2) research papers. For business cards, the reference image is obtained from a high quality upright scan of the card. For museum paintings, we collect data from the Cantor Arts Center at Stanford University for different genres: history, portraits, landscapes and modern-art. The reference images are obtained from the artists' websites like [24] or other online sources. All reference images are high quality JPEG compressed color images. The resolution of reference images varies for each category.

Query Images.

We capture query images with several different camera phones, including some digital cameras. The list of companies and models used is as follows: Apple (iPhone4), Palm (Pre), Nokia (N95, N97, N900, E63, N5800, N86), Motorola (Droid), Canon (G11) and LG (LG300). For product cate-

Data Set	Database (#)	Query (#)	Classes (#)	Rigid	Lighting	Clutter	Perspective	Camera Phone
ZuBuD	1005	115	200	√	–	√	√	–
Oxford	5062	55	17	√	√	√	√	×
INRIA	1491	500	500	–	–	√	√	–
UKY	10200	2550	2550	√	–	–	√	–
ImageNet	11M	15K	15K	–	√	√	√	–
SMVS	1200	3300	1200	√	√	√	√	√

Table 1: Comparison of different data sets. "Classes" refers to the number of distinct objects in the data set. "Rigid" refers to whether on not the objects in the database are rigid. "Lighting" refers to whether or not the query images capture widely varying lighting conditions. "Clutter" refers to whether or not the query images contain foreground/background clutter. "Perspective" refers to whether the data set contains typical perspective distortions. "Camera-phone" refers to whether the images were captured with mobile devices. SMVS is a good data set for mobile visual search applications.

gories like CDs, DVDs, books, text documents and business cards, we capture the images indoors under widely varying lighting conditions over several days. We include foreground and background clutter that would be typically present in the application, e.g., a picture of a CD would might other CDs in the background. For landmarks, we capture images of buildings in San Francisco. We collected query images several months after the reference data was collected. For video clips, the query images were taken from laptop, computer and TV screens to include typical specular distortions. Finally, the paintings were captured at the Cantor Arts Center at Stanford University under controlled lighting conditions typical of museums.

The resolution of the query images varies for each camera phone. We provide the original JPEG compressed high quality color images obtained from the camera. We also provide auxiliary information like phone model number, and GPS location, where applicable. As noted in Tab. 1, the SMVS query data set has the following key characteristics that is lacking in other data sets: rigid objects, widely varying lighting conditions, perspective distortion, foreground and background clutter, realistic ground-truth reference data, and query images from heterogeneous low and high-end camera phones.

Category	Database	Query
CD	100	400
DVD	100	400
Books	100	400
Video Clips	100	400
Landmarks	500	500
Business Cards	100	400
Text documents	100	400
Paintings	100	400

Table 2: Number of query and database images in the SMVS data set for different categories.

Evaluation measures.

A naive retrieval system would match all database images against each query image. Such a brute-force matching scheme provides as an upper-bound on the performance that can be achieved with the feature matching pipeline. Here, we report results for brute-force pairwise matching for different interest point detectors and descriptors using the ratio-test [17] and RANSAC [9]. For RANSAC, we use affine models with a minimum threshold of 10 matches post-RANSAC for declaring a pair of images to be a valid match.

In Fig. 3, we report results for 3 state-of-the-art schemes: (1) SIFT Difference-of-Gaussian (DoG) interest point detector and SIFT descriptor (code: [28]), (2) Hessian-affine interest point detector and SIFT descriptor (code [18]), and (3) Fast Hessian blob interest point detector [2] sped up with integral images, and the recently proposed Compressed Histogram of Gradients (CHoG) descriptor [4]. We report the percentage of images that match, the average number of features and the average number of features that match post-RANSAC for each category.

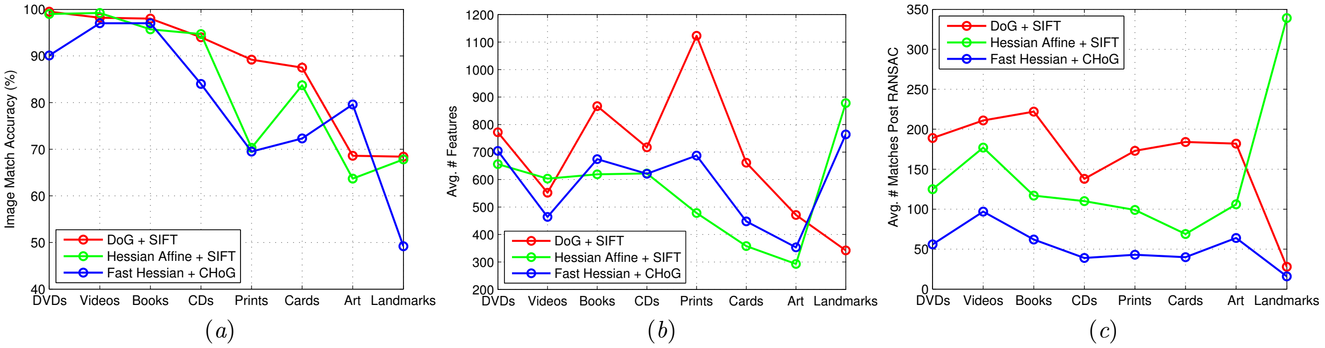
First, we note that indoor categories are easier than outdoor categories. E.g., some categories like CDs, DVDs and book covers achieve over 95% accuracy. The most challenging category is landmarks as the query data is collected several months after the database.

Second, we note that option (1): SIFT interest point detector and descriptor, performs the best. However, option (1) is computationally complex and is not suitable for implementation on mobile devices.

Third, we note that option (3) performs comes close to achieving the performance of (1), with worse performance (10-20% drop) for some categories. The performance hit is incurred due to the fast Hessian-based interest point detector, which is not as robust as the DoG interest point detector. One reason for lower robustness is observed in [26]: the fast box-filtering step causes the interest point detection to lose rotation invariance which affects oriented query images. The CHoG descriptor used in option (3) is a low-bitrate 60-bit descriptor which is shown to perform on par with the 128-dimensional 1024-bit SIFT descriptor using extensive evaluation in [4]. We note that option (3) is most suitable for implementation on mobile devices as the fast hessian interest point detector is an order-of-magnitude faster than SIFT DoG, and the CHoG descriptors generate an order of magnitude less data than SIFT descriptors for efficient transmission [10].

Finally, we list aspects critical for mobile visual search applications. A good image retrieval system should exhibit the follow characteristics when tested on the SMVS dataset.

- High Precision-Recall as size of database increases
- Low retrieval latency
- Fast pre-processing algorithms for improving image qual-



(a) (b) (c)

Figure 3: Results for each category (PR = Post RANSAC). We note that indoor categories like CDs are easier than outdoor categories like landmarks. Books, CD covers, DVD covers and video clips achieve over 95% accuracy.

ity

- Fast and robust interest point detection
- Compact feature data for efficient transmission and storage

4. SUMMARY

We survey popular data sets used in computer vision literature and note that they are limited in many ways. We propose the Stanford Mobile Visual Search data set to overcome several of the limitations in existing data sets. The SMVS data set has several key characteristics lacking in existing data sets: rigid objects, several categories of objects, widely varying lighting conditions, perspective distortion, typical foreground and background clutter, realistic ground-truth reference data, and query data collected from heterogeneous low and high-end camera phones. We hope that this data set will help push research forward in the field of mobile visual search.

5. REFERENCES

[1] Amazon. *SnapTell*, 2007. http://www.snaptell.com.

[2] H. Bay, T. Tuytelaars, and L. V. Gool. SURF: Speeded Up Robust Features. In *Proc. of European Conference on Computer Vision (ECCV)*, Graz, Austria, May 2006.

[3] V. Chandrasekhar, D.M.Chen, S.S.Tsai, N.M.Cheung, H.Chen, G.Takacs, Y.Reznik, R.Vedantham, R.Grzeszczuk, J.Back, and B.Girod. *Stanford Mobile Visual Search Data Set*, 2010. http://mars0.stanford.edu/mvs_images/.

[4] V. Chandrasekhar, G. Takacs, D. M. Chen, S. S. Tsai, R. Grzeszczuk, Y. Reznik, and B. Girod. Compressed Histogram of Gradients: A Low Bitrate Descriptor. In *International Journal of Computer Vision, Special Issue on Mobile Vision*, 2010. under review.

[5] D. M. Chen, N. M. Cheung, S. S. Tsai, V. Chandrasekhar, G. Takacs, R. Vedantham, R. Grzeszczuk, and B. Girod. Dynamic Selection of a Feature-Rich Query Frame for Mobile Video Retrieval. In *Proc. of IEEE International Conference on Image Processing (ICIP)*, Hong Kong, September 2010.

[6] J. Deng, W. Dong, R. Socher, L. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Proc. of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Miami, Florida, June 2009.

[7] B. Erol, E. Antúnez, and J. Hull. Hotpaper: multimedia interaction with paper using mobile phones. In *Proc. of the 16th ACM Multimedia Conference*, New York, NY, USA, 2008.

[8] C. Fellbaum. *WordNet: An Electronic Lexical Database*. Bradford Books, 1998.

[9] M. A. Fischler and R. C. Bolles. Random Sample Consensus: A paradigm for model fitting with applications to image analysis and automated cartography. *Communications of ACM*, 24(6):381–395, 1981.

[10] B. Girod, V. Chandrasekhar, D. M. Chen, N. M. Cheung, R. Grzeszczuk, Y. Reznik, G. Takacs, S. S. Tsai, and R. Vedantham. Mobile Visual Search. In *IEEE Signal Processing Magazine, Special Issue on Mobile Media Search*, 2010. under review.

[11] Google. *Google Goggles*, 2009. http://www.google.com/mobile/goggles/.

[12] L. V. G. H.Shao, T. Svoboda. Zubud-Zürich buildings database for image based recognition. Technical Report 260, ETH Zürich, 2003.

[13] S. J. Matas. Sub-linear indexing for large scale object recognition. In *Proc. of British Machine Vision Conference (BMVC)*, Oxford, UK, June 2005.

[14] H. Jegou, M. Douze, and C. Schmid. Hamming embedding and weak geometric consistency for large scale image search. In *Proc. of European Conference on Computer Vision (ECCV)*, Berlin, Heidelberg, 2008.

[15] Kooaba. *Kooaba*, 2007. http://www.kooaba.com.

[16] Layar. *Layar*, 2010. http://www.layar.com.

[17] D. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004.

[18] K. Mikolajczyk. *Software for computing Hessian-affine interest points and SIFT descriptor*, 2010. http://lear.inrialpes.fr/~jegou/data.php.

[19] MPEG. Requirements for compact descriptors for visual search. In *ISO/IEC JTC1/SC29/WG11/W11531*, Geneva, Switzerland, July 2010.

[20] Newseum. *Newseum*. http://www.newseum.org/todaysfrontpages/hr.asp?fpVname=CA_MSS&ref_pge=gal&b_pge=1.

[21] D. Nistér and H. Stewénius. Scalable recognition with a vocabulary tree. In *Proc. of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, New York, USA, June 2006.

[22] Nokia. *Nokia Point and Find*, 2006. http://www.pointandfind.nokia.com.

[23] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman. Object Retrieval with Large Vocabularies and Fast Spatial Matching. In *Proc. of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Minneapolis, Minnesota, 2007.

[24] W. T. Richards. *William Trot Richards: The Complete Works*. http://www.williamtrostrichards.org/.

[25] J. M. S.Obdrzalek. Image retrieval using local compact dct-based representation. In *Proc. of the 25th DAGM Symposium*, Magdeburg, Germany, September 2003.

[26] G. Takacs, V. Chandrasekhar, H. Chen, D. M. Chen, S. S. Tsai, R. Grzeszczuk, and B. Girod. Permutable Descriptors for Orientation Invariant Matching. In *Proc. of SPIE Workshop on Applications of Digital Image Processing (ADIP)*, San Diego, California, August 2010.

[27] G. Takacs, V. Chandrasekhar, N. Gelfand, Y. Xiong, W. Chen, T. Bismpigiannis, R. Grzeszczuk, K. Pulli, and B. Girod. Outdoors augmented reality on mobile phone using loxel-based visual feature organization. In *Proc. of ACM International Conference on Multimedia Information Retrieval (ACM MIR)*, Vancouver, Canada, October 2008.

[28] A. Vedaldi and B. Fulkerson. VLFeat: An open and portable library of computer vision algorithms, 2008. http://www.vlfeat.org/.

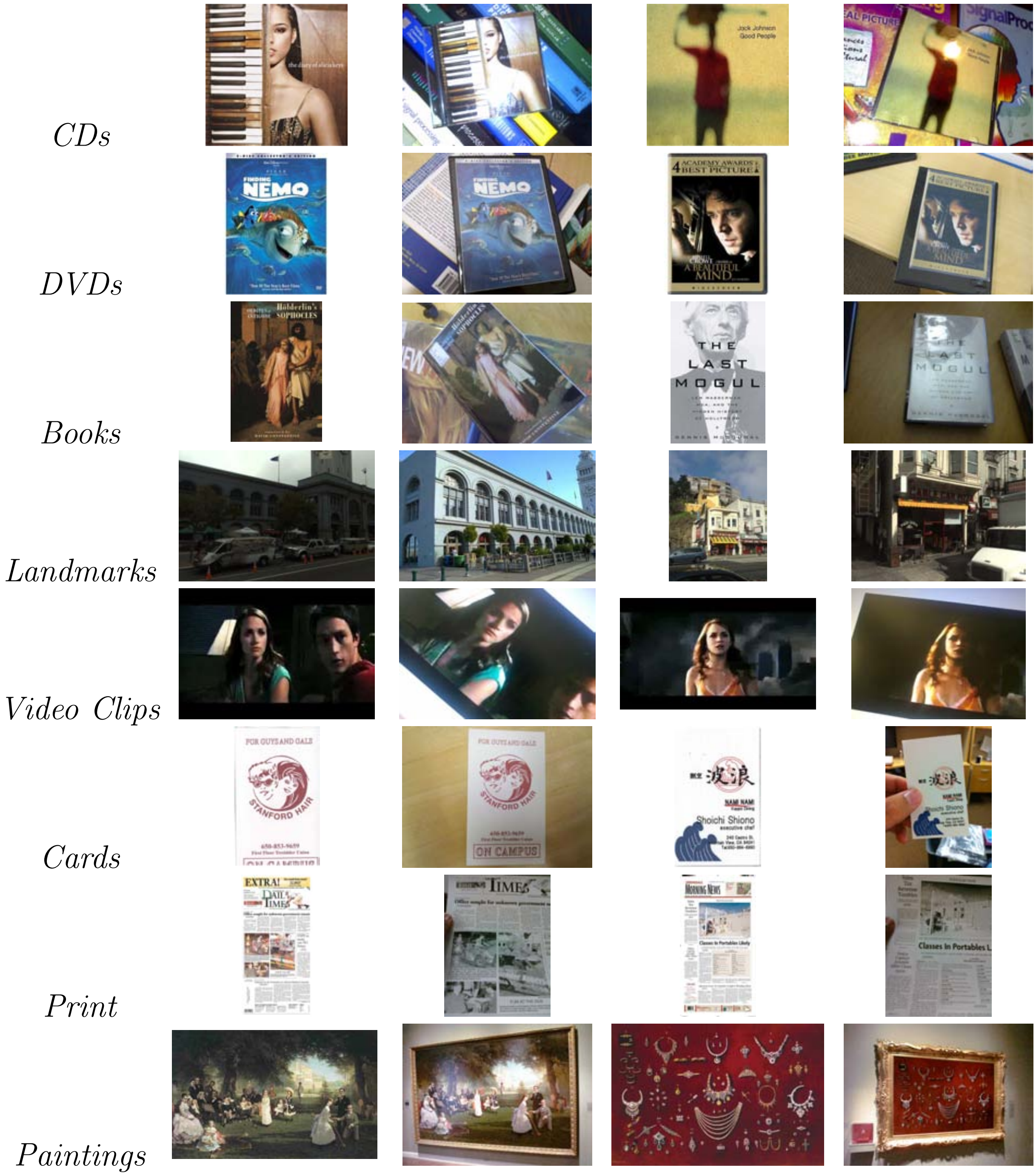


Figure 4: Stanford Mobile Visual Search (SMVS) data set. The data set consists of images for many different categories captured with a variety of camera-phones, and under widely varying lighting conditions. Database and query images alternate in each category.