CALIFORNIA-ND: AN ANNOTATED DATASET FOR NEAR-DUPLICATE DETECTION IN PERSONAL PHOTO COLLECTIONS

Amornched Jinda-Apiraksa, Vassilios Vonikakis, Stefan Winkler

Advanced Digital Sciences Center (ADSC), University of Illinois at Urbana-Champaign, Singapore

{Amornched.ja,bbonik,Stefan.Winkler}@adsc.com.sg

ABSTRACT

Managing photo collections involves a variety of image quality assessment tasks, e.g. the selection of the "best" photos. Detecting near-duplicate images is a prerequisite for automating these tasks. This paper presents a new dataset that may assist researchers in testing algorithms for the detection of near-duplicates in personal photo libraries. The proposed dataset is derived directly from an actual personal travel photo collection. It contains many difficult cases and types of near-duplicates. More importantly, in order to deal with the inevitable ambiguity that the near-duplicate cases exhibit, the dataset is annotated by 10 different subjects. These annotations are combined into a non-binary ground truth, which indicates the probability that a pair of images may be considered a near-duplicate by an observer.

Index Terms— Photo quality, photowork, near-duplicate images, user study, annotation

1. INTRODUCTION

Most people nowadays have at least one digital camera with them at all times, mainly due to the widespread use of smart phones. Additionally, the affordability of digital images has made it very common for camera users to grab more than one picture of the same scene, in order to increase the chances of having a good-quality shot [1]. This has led to a constant increase in photo library size for the average user and has introduced a new important problem: photo libraries are cluttered with images that are slightly different, but depict the same or almost the same scene. These images – generally known as "near-duplicates" (NDs) – have a negative impact not only on the size of photo libraries, but generally the quality of the photo managing and browsing experience.

Detecting NDs in a set of images is an important step in the photowork process [1]. Users have to manually go through the set of images, identify ND cases and then usually keep the one(s) with the highest image quality. Thus, ND detection can be considered a prerequisite stage to the actual comparison of photos, by confining the image quality assessment step to a specific subset and not the whole photolibrary.

According to [2], ND cases can be grouped into two categories: identical ND (IND) images, which are derived from the same digital source after applying some transformations, and non-identical ND (NIND) images, which share the same scenes or objects. The subjectivity in interpretation that characterizes NIND images has resulted in the majority of the existing work to focus mainly on the simpler case of IND, which is common in the domains of copyright detection or duplicate search in the web. This has a profound effect on the available datasets used for testing ND detection algorithms. Table 1 shows the most commonly used datasets. Most of them comprise frames taken from news clips, movies, sports events, buildings, objects etc. In many of them, artificial degradations are applied to the original set of images, like cropping, blurring, or other kinds of filtering, in order to create variations of the originals, with the latter serving as ground truth (GT). This inevitably leads to a binary decision; a test image A is considered to be an IND of another image B if it is derived from B through the use of transformations.

This approach however is not adequate for personal photo collections, which mostly comprise NIND cases. There are two reasons for this. First, personal photo collections usually contain travel photos, lots of portraits, family or group photos in various activities and scenery, which may be quite different from the photos used in IND datasets. Second, binary decisions are not suitable for NIND cases, since there is a considerable degree of subjectivity in interpretation, as the examples in Fig. 1 indicate. This is due mainly to the semantic gap that may result in different interpretations between observers, and which binary ground truth cannot adequately capture. This subjectivity has discouraged researchers from working with NIND cases. Many other papers have explicitly mentioned this issue before:

- *"Deciding if two images are duplicates is highly subjective. In addition, when there is 100% agreement duplicate images can be visually different and non-duplicate images can be visually very similar."* [28]
- *"We do not have access to ground-truth data for our experiments, since we are not aware of any large pub-*

This study is supported by the research grant for ADSC's Human Sixth Sense Programme from Singapore's Agency for Science, Technology and Research (A*STAR).

Table 1. Overview of existing datasets.

Database	Content	Annotation/Rating	Used in
TRECVID [3]	Frames of news videos from various TV stations	Many annotations available	[4–10]
MUSCLE VCD [11]	Documentaries, sports events, movies, TV shows, cartoons	Ground truth annotations for 15 queries (transformed videos)	[12, 13]
Google search engine web crawling	Images of various objects and people	N/A	[14–17]
INRIA Copydays dataset [18, 19]	Personal holiday photos with artificial degradations (no people)	500 queries and their correct retrieval results	[12, 20]
Oxford buildings dataset [21, 22]	5062 images of landmarks collected from Flickr	Ground truth for 11 different landmarks, with 5 possible queries	[15, 20]
Internet partial-duplicate image database [23]	Brand logos	N/A	[24]
UKbench dataset (object recognition evaluation) [25]	2550 groups of 4 images, from four different viewpoints.	N/A	[20]
Corel Photo CD collection	Scenery, animals, flowers, object close-ups, activities	N/A	[4, 17]
Flickr	Various images used as distractions for other databases	N/A	[12, 20, 26, 27]

lic corpus in which near duplicate images have been annotated." [5]

- *"Labeling of large data sets is difficult in its own right and the subjective definition of near duplicate images complicates things further."* [8]

The proposed dataset attempts to bridge this gap; it is designed specifically to tackle the problem of subjectivity in the interpretation of NIND cases in personal photo collections. It comprises images taken during a vacation in California (hence the name of the dataset). The majority of images are in exactly the same sequence as they were captured. There are many different ND cases, ranging from the typical ones (zooming, panning etc.) to others that may be less common, such as panorama shots, performance images, or burst shots. More importantly, in order to deal with the inevitable ambiguity of the ND detection process, the dataset is annotated by 10 different subjects, including the photographer. These annotations are combined into a non-binary ground truth, which indicates the probability that a pair of images may be considered a ND by an observer. Researchers may use the proposed dataset to evaluate their ND detection algorithms or to study the correlation between subjectivity in ND cases and specific image features.

The paper is organized as follows. Section 2 describes the proposed dataset in detail. Section 3 discusses the user study for the annotation of the photos and also presents some statistics derived from the analysis of the ground truth. Section 4 provides an overview of the specific content included with the California-ND dataset. Some concluding remarks are given in Section 5.

2. DATASET DESCRIPTION

The dataset comprises 701 photos from an actual user's photo collection documenting a holiday trip, which roughly coincides with the average number of photos taken per trip [29]. The first 604 photos were *consecutively* selected from the beginning of the collection. The remaining 97 photos consist of pairs or groups of interesting cases from the same collection that are not present among the first 604 photos. Completely manual selection of all photos in the dataset would result in a bias regarding the time stamps of the photos, making it easier to identify ND by just clustering the time in which the photos were captured. Fig. 1 depicts some sample NIND cases from the dataset.

The included ND cases are the three basic ones reported in [10, 28], i.e. variations in scene, camera, and image. *Scene variations* come from changes or movements on target objects or background. *Camera variations* occur when the camera's settings were changed. *Image variations* range from noise, color saturation, resolution etc. All of them may af-

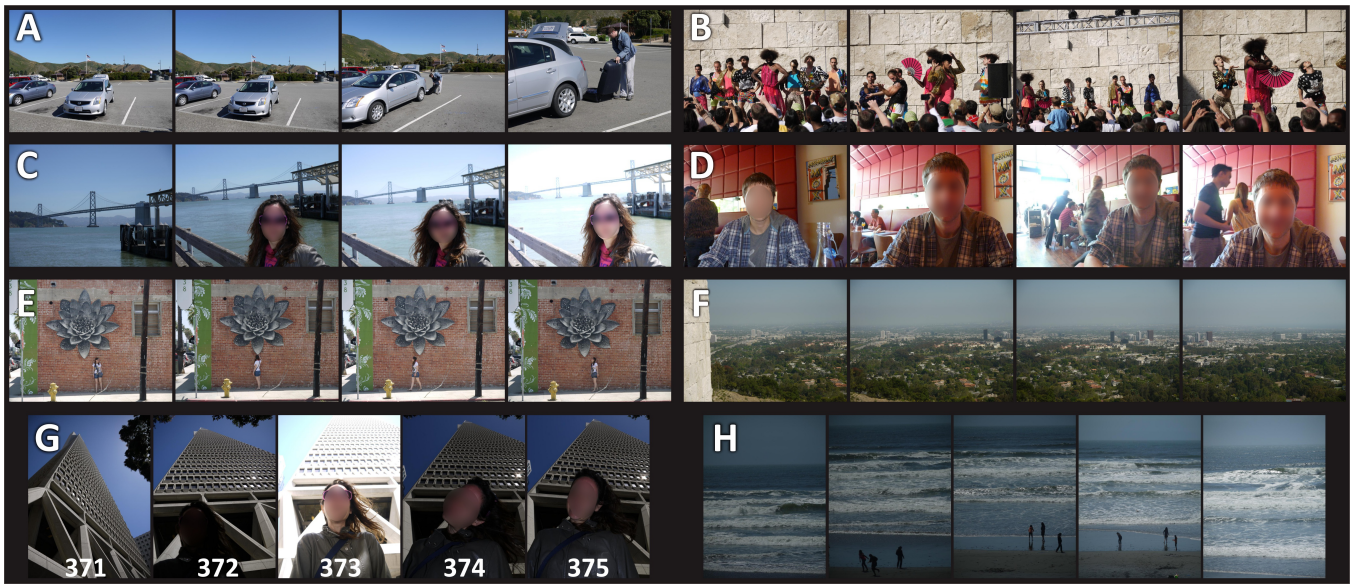

Fig. 1. Sample images from the California-ND dataset. A: Severe zooming and viewpoint change. B: Performance shots. C: Different exposures. D: Viewpoint and background change. E: Burst shots. F: Panorama. G & H: Combination of changes.

fect photo quality. Some other examples include group photos with varying numbers of people, orientation differences, portrait photos in front of the crowd, road side photos, as well as similar photos from two different cameras with unsynchronized time-stamps. A more detailed description of the included ND cases follows.

1. **Burst shots** (344 photos, 49%) are sets of photos taken consecutively in a very short period of time, resulting in a high degree of visual similarity. There are 101 such sets, with 2 to 12 photos per set.

2. **Moving Background shots** (149 photos, 21%) are *burst shots* which contain the same foreground object(s), but exhibit considerable differences in the background due to the nature of the scene, e.g. cars or people passing by.

3. **Show/Performance shots** (54 photos, 8%) are sets of photos taken during a show or performance. Those photos are not necessarily visually similar due to the amount and variety of movement in the show. However, it is obvious to humans that they have the same theme. For browsing purposes, it would be more convenient to group these photos together. There are 7 different performance sets, with 2 to 23 photos each.

4. **Group photos** (17 photos, 2%) have more than one target person. There might be slight movements or pose changes between shots. However, it is very obvious to humans that they were taken at the same time and place. In the case of group photos with many people, there might be different numbers of people in some of the photos.

5. **Panorama shots** (8 photos, 1%) are sets of photos taken by continuously panning the camera in one direction, with the intention of merging the photos into a panoramic view. Typical panorama shots have about 30% overlapped area [30]. Grouping these kind of photos generally helps to reduce selection and browsing time.

6. **Exposure/Brightness Difference** (58 photos, 8%) can be caused by many factors, e.g. different exposure settings, shutter speed, ISO, flash, or even environmental changes such as clouds covering the sun.

7. **Viewpoint Difference/Zooming** (36 photos, 5%) includes photos taken from different points of view or with extended zooming on the scene. Note that zooming is usually different from a simple cropping that is used in existing datasets, since by the time the camera lens has zoomed and focused, the scene may also have changed.

8. **Focus Change** (22 photos, 3%) includes photos of the same scene with different focusing points or depth of field. The focus change might be intentional (emphasize different objects) or accidental (camera cannot or missed focus).

9. **White Balance Difference** (8 photos, 1%) may be due to either the user choosing different settings manually

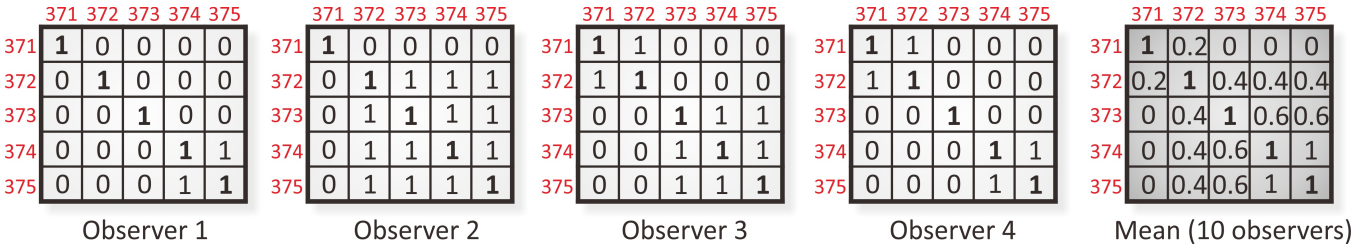

Fig. 2. A small section of the correlation matrices for the 5 photos of Fig. 1G. The binary ratings for 4 observers are shown, along with the average over all 10 subjects.

or the automatic selection of different white balance algorithms by the camera. As a result, the same scene may appear to have different dominant tint in different photos.

The cases of burst, performance, and panorama shots have not been considered in any other studies before. Since the dataset comprises real and not synthetic photos, many cases have a combination of degradations. As a result, the above percentages do not add up to 100%.

3. GROUND TRUTH

10 subjects (9 male, 1 female, with ages ranging between 23-40 years old) were presented with the 701 photos of the dataset. One of them was the photographer of the collection. All subjects had some experience with digital photography (they had their own digital cameras, taking pictures while traveling or during family moments), but no particular specialization in the art of photography. As a result, they may be considered "average users". The subjects were asked to go through all of the photos of the dataset, freely, with no time constraint, using a browsing software of their own preference, and with no restrictions regarding the browsing method (consecutive/random), while trying to identify groups of ND photos. On average, the whole process lasted between 30-45 minutes for each subject. The preferred browsing software was Windows Explorer (with large or extra large icon settings), along with Windows Photo Viewer, while the images were mainly viewed consecutively.

Regarding the definition of ND, subjects were specifically advised that there is no right or wrong answer. The following three guidelines were provided in order to assist them in this task, explicitly mentioning that not all of them should necessarily hold at the same time:

1. *"If any two (or more) images look similar in visual appearance, or convey similar concepts to you, label them as near-duplicates."*

2. *"Near-duplicates are photos you would not want to see more than once when browsing the collection."*

3. *"Near-duplicates are photos which if grouped together can save you time when browsing the collection."*

If the subjects thought that any pair (or group) of images were ND, they listed the photo ID numbers in a text file.

Subsequently, the groupings provided by each subject were converted into a correlation matrix of size is 701×701, containing all possible pair combinations of the photos in the dataset; its cell (y,x) has a value of 1 if the image pair (y,x) is considered a ND case, and 0 otherwise. Since every image can be considered to be a ND of itself, the diagonal of this matrix is always 1. The matrix is also symmetric, since near-duplicate pairs are commutative, and sparse, since most random image pairs are not related.

Obviously not all subjects fully agreed on the same photos. However, as mentioned before, there is no right or wrong in this task. Therefore, the binary correlation matrices from the 10 subjects were averaged, resulting in a non-binary correlation matrix with values in the interval [0,1]. This value reflects the agreement between subject, or in other words, the probability that a pair of images is considered ND by observers.

Fig. 2 depicts a small part of the correlation matrices of the ratings of 4 subjects, as well as the GT correlation matrix, for the 5 images of Fig. 1G. These binary correlation matrices are very different for each observer, clearly highlighting the subjectivity of the task. For example, images 374 and 375 were considered ND by all subjects. On the other hand, images 371 and 372 are an ambiguous case, which was considered to be ND by only 20% of the subjects. While there are clear differences between subjects, we could not find any specific aspect of the photographer's annotations that stood out from the other 9 participants.

There are a total of 245350 unique possible combinations of image pairs in the correlation matrix (pairs AB and BA are the same) for all 701 photos in the dataset. The majority of image pairs (240741) are unrelated, and all subjects agreed that they are not ND. There are 4609 image pairs which at least one subject identified as ND. The distribution of the level of agreement between subjects for these pairs is shown in Fig. 3. Only in 18% of these cases all subjects agreed, whereas in 82% of cases subjects disagreed to some extent whether or not a pair of images should be considered ND.

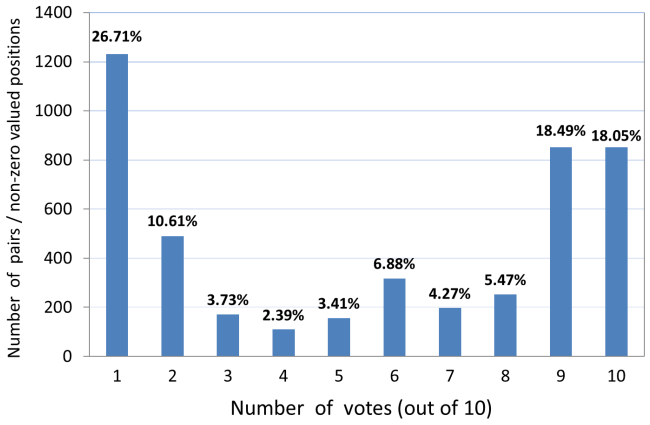

Fig. 3. The distribution of the agreement on pairs of photos among all 10 subjects. Since there are too many non ND pairs, the number of zero votes was deliberately omitted, for visualization purposes.

Furthermore, the distribution of Fig. 3 is almost balanced: the sum of percentages for votes between 1-5 (46.85%), which indicates disagreement between subjects, is approximately equal to the votes between 6-10 (53.15%), which indicates majority agreement. This clearly demonstrates the subjective nature of ND detection in personal photo collections, with subjects applying different criteria and/or different semantic interpretations. Consequently, any purely binary ground truth will fail to capture the subjectivity of this task. The proposed dataset, accompanied with the non-binary GT, offers an alternative approach which reflects the subjective nature of NIND detection.

4. DOWNLOAD AND LICENSE

Although all photos are made freely available to public, the privacy of the owner has to be preserved. Therefore, the two main subjects' faces in the dataset were blurred with a Gaussian filter, which concerns 182 photos (about a quarter of the total).[1] However, since face detection/recognition may be part of a ND detection algorithm, the relevant information (bounding boxes of the blurred faces and subject identifiers) is provided in a separate file.

The proposed California-ND dataset is made available under a Creative Commons License and can be downloaded from http://vintage.winklerbros.net/californiaND.html. It comprises:

- 701 photos (resolution 1024×768) with original embedded EXIF data from the cameras.
- The near-duplicate annotations of 10 subjects, including the photographer.
- The resulting binary correlation matrices for each of the 10 subjects.
- The final GT correlation matrix with the average ratings from all subjects.
- A listing of identifier and location (bounding boxes) of all blurred faces.

[1] In order to ensure that the blurring of faces does not have a significant impact on the detection of near duplicates, we tested the performance of an existing ND detection system [31] on the original and the blurred-faces versions of the dataset. The resulting performance difference is less than 1% (naturally, this also depends on the specific choice of features).

5. CONCLUSION

We have presented a new dataset for the detection of non-identical near-duplicate (NIND) images in personal photo collections. Contrary to other existing datasets in the same domain, it is taken directly from a real user's photo collection, maintaining the original image sequence as much as possible. It includes many challenging NIND cases without the use of artificial transformations. Another unique feature is that it includes non-binary ground truth, which has been constructed by averaging the individual binary annotations of 10 different subjects, thus representing the probability that a particular image pair may be considered a near-duplicate case by a user. This approach is more in line with the subjective nature of the NIND detection task. Researchers may use the proposed dataset to evaluate their ND detection algorithms or to study the correlation between subjectivity in ND cases and specific image features.

6. REFERENCES

[1] D. Kirk, A. Sellen, C. Rother, and K. Wood, "Understanding photowork," in *Proc. SIGCHI*, Montreal, Canada, April 2006, pp. 761–770.

[2] J. J. Foo, R. Sinha, and J. Zobel, "Discovery of image versions in large collections," in *Proc. ACM Multimedia Modeling*, Singapore, Jan. 2007.

[3] Paul Over, "TREC video retrieval evaluation: TRECVID," http://trecvid.nist.gov/, Jan. 2013.

[4] J. J. Foo and R. Sinha, "Using redundant bit vectors for near-duplicate image detection," in *Proc. DASFAA*, Bangkok, Thailand, April 2007, pp. 472–484.

[5] O. Chum, J. Philbin, M. Isard, and A. Zisserman, "Scalable near identical image and shot detection," in *Proc. CIVR*, Amsterdam, The Netherlands, July 2007, pp. 549–556.

[6] J. Zhu, S. C. H. Hoi, M. R. Lyu, and S. Yan, "Near-duplicate keyframe retrieval by semi-supervised learning and nonrigid image matching," *ACM Trans. Multimedia Comput. Commun. Appl.*, vol. 7, no. 1, pp. 4:1–4:24, Feb. 2011.

[7] D. Xu, T. J. Cham, S. Yan, L. Duan, and S. F. Chang, "Near duplicate identification with spatially aligned pyramid matching," *IEEE Trans. Cir. and Sys. for Video Technol.*, vol. 20, no. 8, pp. 1068–1079, Aug. 2010.

[8] O. Chum, J. Philbin, and A. Zisserman, "Near duplicate image detection: min-Hash and tf-idf weighting," in *Proc. BMVC*, Leeds, UK, Sept. 2008.

[9] Y. Wang, Z. Hou, and K. Leman, "Keypoint-based near-duplicate images detection using affine invariant feature and color matching," in *Proc. ICASSP*, Prague, Czech Republic, May 2011, pp. 1209–1212.

[10] Dong-Qing Zhang and Shih-Fu Chang, "Detecting image near-duplicate by stochastic attributed relational graph matching with learning," in *Proc. ACM Multimedia*, New York, USA, Oct. 2004, pp. 877–884.

[11] INRIA, "Video copy detection evaluation showcase," https://www.rocq.inria.fr/imedia/civr-bench/data.html, 2007.

[12] Ligang Zheng, Guoping Qiu, Jiwu Huang, and Hao Fu, "Salient covariance for near-duplicate image and video detection," in *Proc. ICIP*, Brussels, Belgium, Sept. 2011, pp. 2537–2540.

[13] X. Yang, Q. Zhu, and K. T. Cheng, "Near-duplicate detection for images and videos," in *Proc. ACM LS-MMRM Workshop*, Beijing, China, Oct. 2009, pp. 73–80.

[14] J. J. Foo, R. Sinha, and J. Zobel, "SICO: A system for detection of near-duplicate images during search," in *Proc. ICME*, Beijing, China, July 2007, pp. 595 –598.

[15] D. C. Lee, Q. Ke, and M. Isard, "Partition min-hash for partial duplicate image discovery," in *Proc. ECCV*, Heraklion, Greece, Sept. 2010, pp. 648–662.

[16] Z. Wu, Q. Ke, M. Isard, and J. Sun, "Bundling features for large scale partial-duplicate web image search," in *Proc. CVPR*, Florida, USA, June 2009, pp. 25–32.

[17] J. J. Foo, J. Zobel, and R. Sinha, "Clustering near-duplicate images in large collections," in *Proc. ACM MIR*, Augsburg, Germany, Sept. 2007, pp. 21–30.

[18] H. Jégou, M. Douze, and C. Schmid, "Hamming embedding and weak geometric consistency for large scale image search," in *Proc. ECCV*, Marseille, France, Oct. 2008, vol. 1, pp. 304–317.

[19] H. Jégou, M. Douze, and C. Schmid, "INRIA holidays dataset," http://lear.inrialpes.fr/~jegou/data.php, Oct. 2008.

[20] H. Xie, K. Gao, Y. Zhang, S. Tang, J. Li, and Y. Liu, "Efficient feature detection and effective post-verification for large scale near-duplicate image search," *IEEE Trans. Multimedia*, vol. 13, no. 6, pp. 1319 –1332, Dec. 2011.

[21] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman, "Object retrieval with large vocabularies and fast spatial matching," in *Proc. CVPR*, Minneapolis, MN, June 2007.

[22] J. Philbin, R. Arandjelović, and A. Zisserman, "The Oxford building dataset," http://www.robots.ox.ac.uk/~vgg/data/oxbuildings/, June 2007.

[23] Z. Wu, Q. Xu, S. Jiang, Q. Huang, P. Cui, and L. Li, "Adding affine invariant geometric constraint for partial-duplicate image retrieval," in *Proc. ICPR*, Istanbul, Turkey, Aug. 2010, pp. 842 –845.

[24] L. Li, Z. Wu, Z. J. Zha, S. Jiang, and Q. Huang, "Matching content-based saliency regions for partial-duplicate image retrieval," in *Proc. ICME*, Barcelona, Spain, July 2011, pp. 1–6.

[25] D. Nistér and H. Stewénius, "Scalable recognition with a vocabulary tree," in *Proc. CVPR*, New York, NY, June 2006, vol. 2, pp. 2161–2168.

[26] Wei-Ta Chu and Chia-Hung Lin, "Consumer photo management and browsing facilitated by near-duplicate detection with feature filtering," *Vis. Comm. Image Repres.*, vol. 21, no. 3, pp. 256–268, April 2010.

[27] H. S. Kim, H. W. Chang, J. Lee, and D. Lee, "BASIL: Effective near-duplicate image detection using gene sequence alignment," in *Proc. ECIR*, Milton Keynes, UK, March 2010, pp. 229–240.

[28] A. Jaimes, S. Chang, and A.C. Loui, "Detection of non-identical duplicate consumer photographs," in *Proc. ICICS & PCM*, Singapore, Dec. 2003, pp. 16–20.

[29] A. Loos, R. Paduschek, and D. Kormann, "Evaluation of algorithms for the summarization of photo collections," in *Proc. Theseus/ImageCLEF Workshop on Visual Information Retrieval Evaluation*, Corfu, Greece, Sept. 2009.

[30] C. Jacobs, *Interactive Panoramas: Techniques For Digital Panoramic Photography*, X.Media.Publishing Series. Springer, Nov. 2004.

[31] V. Vonikakis and S. Winkler, "Emotion-based sequence of family photos," in *Proc. ACM Multimedia*, Nara, Japan, Nov. 2012.