PhotoCluster
A Multi-clustering Technique for Near-duplicate Detection in Personal Photo Collections

Vassilios Vonikakis, Amornched Jinda-Apiraksa and Stefan Winkler
Advanced Digital Sciences Center (ADSC), University of Illinois at Urbana-Champaign, Singapore, Singapore

Keywords: Near Duplicate Detection, Image Similarity, Personal Photo Collections.

Abstract: This paper presents PhotoCluster, a new technique for identifying non-identical near-duplicate images in personal photo collections. Contrary to existing methods, PhotoCluster estimates the probability that a pair of images may be considered near-duplicate. Its main thrust is a multiple clustering step that produces a non-binary near-duplicate probability for each image pair, which exhibits correlation with the average observer opinion. First, PhotoCluster partitions the photolibrary into groups of semantically similar photos, using global features. Then, the multiple clustering step is applied within the images of these groups, using a combination of global and local features. Computationally expensive comparisons between local features are taking place only on a limited part of the library, resulting in a low overall computational cost. Evaluation with two publicly available datasets show that PhotoCluster outperforms existing methods, especially in identifying ambiguous near-duplicate cases.

1 INTRODUCTION

It is common nowadays for people to carry at least one digital camera with them, mainly due to the widespread use of smart phones. Additionally, the affordability of digital images allows camera users to easily take more than one picture of the same scene, in order to increase the chances of capturing an appealing photo. This has introduced an important new problem: photolibraries are cluttered with near-duplicate (ND) images that are similar and thus redundant. This negatively affects not only the size of photolibraries, but also various photowork tasks, such as triaging (Kim et al., 2012) or browsing.

According to Foo et al. (2007a), ND cases can be grouped into two categories: identical ND (IND), which are derived from the same digital source after applying some transformations, and non-identical ND (NIND), which are images of the same scene or objects. Personal photolibraries may comprise a high number of NIND cases. Identifying these cases is challenging, since they exhibit a considerable degree of subjectivity in interpretation, as Fig. 1 indicates. According to Jinda-Apiraksa et al. (2013), in only 20% of images taken from personal photo-collections, do observers completely agree that a pair may be ND. This clearly demonstrates that ND detection in personal photolibraries is a very challenging task, mainly due to the semantic gap, which may result in different interpretations between observers. Due to this subjectivity, most existing works have focused on IND cases, such as copyright detection or duplicate search on the web (Wu et al., 2009), which have a more objective definition.

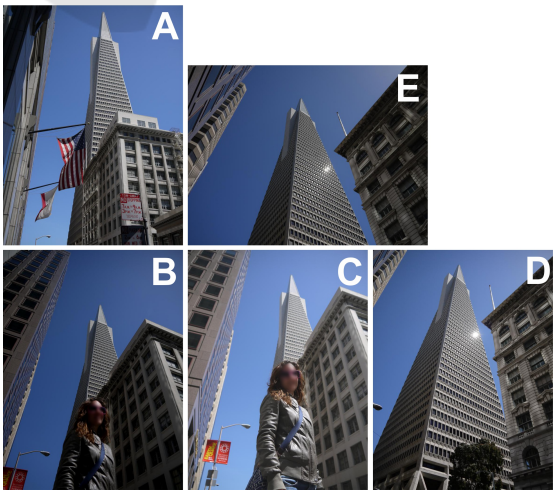

Figure 1: A typical case of NIND. Observers may disagree which of the five images should be considered ND.

The first attempts to tackle the problem of ND

Vonikakis V., Jinda-Apiraksa A. and Winkler S..
PhotoCluster - A Multi-clustering Technique for Near-duplicate Detection in Personal Photo Collections.
DOI: 10.5220/0004737701530161
In *Proceedings of the 9th International Conference on Computer Vision Theory and Applications* (VISAPP-2014), pages 153-161
ISBN: 978-989-758-004-8

were extensions of image retrieval (Zhang and Chang, 2004). Since ND detection is computationally intensive, many methods have proposed a cascade structure in order to quickly exclude images that are clearly not ND, in order to limit the most computationally intensive processing to as few images as possible (Tang and Gao, 2009; Wang et al., 2011). Various types of hashing have been used for this purpose, such as locality sensitive hashing (Foo et al., 2007a; Ke et al., 2004), as has saliency (Zheng et al., 2011). Recently, clustering methods have emerged as a promising approach to the ND problem (Chu and Lin, 2010; Foo et al., 2007b; Wang et al., 2012; Zhao and Ngo, 2009).

Although there is an extensive body of work regarding ND images, very few techniques are specifically designed for personal photo collections. Notable exceptions are Jaimes et al. (2003, 2002); Tang and Gao (2009). However, they are based on binary decisions, which cannot capture the subjectivity of personal photo collections (Jinda-Apiraksa et al., 2013).

PhotoCluster specifically addresses the subjectivity of NIND cases in personal photo collections. It automatically partitions the photolibrary into smaller sets of similar images, on which a multiple clustering step is applied. In every iteration, the parameter controlling the number and size of the clusters varies, resulting in different numbers and types of clusters. These multiple binary results are combined into a correlation matrix with non-binary entries. As a result, contrary to most existing approaches, PhotoCluster produces continuous values, which exhibit correlation with the probability that a pair of images may be considered ND by observers. Since the F1 score, which is the main comparison metric used in these tasks, cannot be directly applied to non-binary values, we generalize the formulas for the calculation of True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN). This results in a *soft* version of the F1 score, which can be applied to continuous as well as binary values. Evaluations with two different datasets show that PhotoCluster represents a promising approach for tackling the subjectivity of NIND in personal photo collections, outperforming existing techniques in the detection of ambiguous ND cases.

The rest of the paper is organized as follows. Section 2 describes the proposed method and its implementation. Section 3 describes the new generalized formulas for the soft F1 score. Experimental results are reported in Section 4, and concluding remarks are presented in Section 5.

2 PHOTOCLUSTER

2.1 Method

The block diagram of the PhotoCluster method is depicted in Fig. 2. The first stage involves the extraction of global features from the whole photolibrary, which is usually a lot less computationally intensive than the extraction of local ones. Based on these global features, and using image dissimilarity (Vonikakis and Winkler, 2012) as a distance metric, the photolibrary *PL*, comprising N number of images, is partitioned into L image sets, using the Affinity Propagation (AP) clustering technique (Frey and Dueck, 2007), which automatically determines the number of clusters. Local features are extracted and matched only within the images in each of these L sets. This organization strategy is selected in order to confine the computationally intensive detection/matching process to only a small set of similar images (based on global features) and not to the whole photolibrary. Additionally, the resulting L image sets usually contain semantically similar images, which can be useful for image browsing or summarization.

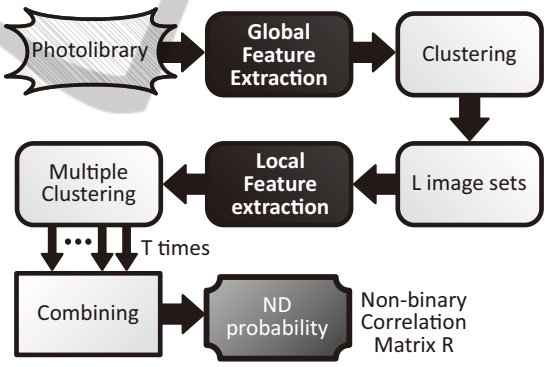

Figure 2: Block diagram of the PhotoCluster method.

For each of these L sets, a multiple clustering step is applied independently, based on a distance metric combining both local and global features. Let $S = \{I_1, ..., I_n\}$ be one of the L sets, comprising n images. Let $D_S = [d_{ij}]_{n \times n}$ be the dissimilarity matrix of all possible image pairs in S, with d_{ij} denoting the dissimilarity between images I_i and I_j. Since $d_{ij} = d_{ji}$, matrix D_S is symmetric. Let F be a clustering method applied on the dissimilarity matrix D_S, such that $F(D_S, k) = \{C_1, ..., C_m\}$, partitioning S into a set of m image clusters, such that $C_1 \cup ... \cup C_m = S$, with k being the parameter that controls m. The proposed multiple clustering step controls k in such way that different clusters will be formed with every iteration. For clustering methods that directly control m, like K-means, k equals m and can be adjusted in the

range $[k_1, k_q]$, with $F(D_S, k_1) = S$ (all images are in one cluster) and $F(D_S, k_q) = \{C_1, ..., C_n\}$ (every image is an individual cluster). For clustering methods in which there is no direct control over m, like AP, k represents the strictness level of F and should be chosen within its allowed limits (for AP, k represents the *preference* parameter, with $k_1 = 0$ and $k_q = \max[D_S]$).

For every image pair u_{ij}^S of images I_i and I_j, the probability P for it to be considered ND is given by:

$$P\left(u_{ij}^S\right) = \frac{\sum_{x=1}^{q}\left[h_x(u_{ij}^S)w(k_x)\right]}{\sum_{x=1}^{q} w(k_x)}, \quad (1)$$

$$h_x\left(u_{ij}^S\right) = \begin{cases} 0 & \text{if } u_{ij}^S \notin Q_x \\ 1 & \text{if } u_{ij}^S \in Q_x \end{cases}, \quad (2)$$

$$w(k) = a\left(\frac{k-k_1}{k_q-k_1}\right) + (1-a)\left(\frac{k_q-k}{k_q-k_1}\right), \quad (3)$$

where $Q_x = \{\mathcal{P}(C_1), ..., \mathcal{P}(C_x)\}$ is the set of powersets of each member of $F(D_S, k_x) = \{C_1, ..., C_x\}$, $h_x(u_{ij}^S)$ is a function that outputs 1 when the images of the pair u_{ij}^S are grouped in the same cluster during the x^{th} iteration, $w(k_x)$ is its weighting factor, and a is an estimation parameter related to the average degree of dissimilarity between all images of S, controlling the contribution of $h_x(u_{ij}^S)$. As such, a is the mean value of all entries of D_S, with $a \in [0, 1]$.

Fig. 3 depicts the graphical representation of the weighting function used, for different values of a. In the case that all images of set S are identical (zero dissimilarity), $a = 0$ and thus greater importance will be given to the clustering iterations with parameter k close to k_1, which tend to group all images in one cluster. In the opposite case, if all images of set S are totally dissimilar (maximum dissimilarity), $a = 1$ and thus greater importance will be given to the clustering iterations with parameter k close to k_q, which tend to keep each image in a separate group. Any other case between these two extremes will be a linear combination.

The intuition behind the proposed approach is that, although the number or size of clusters may change as parameter k changes with each iteration, stronger ND cases (with a small dissimilarity d) will be in the same cluster most of the time, resulting in a higher probability value. On the other hand, ambiguous ND cases (with larger dissimilarity d), will tend to be clustered together less often, resulting in a lower probability value.

Since the multiple clustering step is applied only on images of the same set S, the probability that an image A from set S_i is ND with an image B from another set S_j is 0. The final output of the method is a correlation matrix $R = [P_{ij}]_{N\times N}$ containing the ND probability for every possible image pair of PL. The non-binary probabilistic output of PhotoCluster can be directly used in cases such as spatial positioning of similar images for photo browsing (Schaefer, 2010) or the automatic selection of images for slideshow creation (Vonikakis and Winkler, 2012). In cases where a binary decision is needed, a final optional step could be included for the binarization of R. Since this is a highly subjective process, some user control is necessary, e.g. by setting a personalized probability threshold. Alternatively, other more sophisticated methods could be used, taking into consideration prior user activity and preferences.

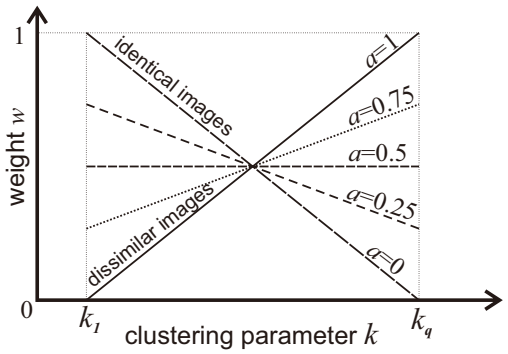

Figure 3: Weighting function w (Eq. 3) for different values of a.

2.2 Implementation

The image distance metric used for the implementation of PhotoCluster is given by the following formula from Vonikakis and Winkler (2012):

$$d_{AB} = 1 - \frac{\sum_{x=1}^{Z}(1 - f_{AB}^x)W^x}{\sum_{x=1}^{Z} W^x}, \quad (4)$$
$$d_{AB}, f_{AB}^x \in [0,1], W^x \in \mathbb{R}^+,$$

where d_{AB} is the final dissimilarity between images A and B. Z is the total number of features, f_{AB}^x is the x^{th} feature distance, normalized to the interval $[0, 1]$, and W^x its importance weight. Note that since similarity and dissimilarity are two complementary concepts, the former can be expressed as $1 - d_{AB}$.

The multiple clustering step of PhotoCluster was implemented using AP (Frey and Dueck, 2007), in which we control the *preference* parameter that indirectly adjusts the number and size of clusters, ranging from zero ($k_1 = 0$) to the maximum dissimilarity values of every set S ($k_q = \max[D_S]$). When $k_1 = 0$ AP tends to group all images into one cluster. On the other hand, when $k_q = \max[D_S]$, every image tends to be an individual cluster.

$Z = 2$ global features are used to partition the photolibrary into L sets, namely color histograms and time stamps, with equal importance weights. $Z = 3$ features are used in the multiple clustering step, namely SIFT Matching Ratio (SMR), color histograms, and timestamps, in descending order of importance. SMR between two images A and B is given by the following equation:

$$SMR_{AB} = \frac{2M_{AB}}{(K_A + K_B)}, \quad (5)$$

where K denotes the number of keypoints from each image, and M_{AB} is the number of bidirectional matches between the two images.

It should be noted that additional features could easily be included in the distance metric of Eq. 4. For example, GIST (Oliva and Torralba, 2001) or geotagging data could be added in the global features used for the initial partitioning of the photolibrary, whereas other, computationally more expensive features such as co-saliency (Fu et al., 2013) could be included in the multiple clustering step.

3 SOFT F1 SCORE

The most common metrics used for the evaluation of ND results by almost all methods are precision, recall, as well as the F1 score, which is a combination of both. These metrics however rely on binary decisions for the calculation of True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN). This only applies if we assume that a pair of images can either be ND or not; no in-between condition can exist. Since this assumption does not hold in the case of NIND, we introduce an extension of the classic F1 score, the soft F1 score (sF1), which can be applied to both binary and continuous values.

Let $R = [r_{ij}]_{N\times N}$ be the correlation matrix of the results, and $G = [g_{ij}]_{N\times N}$ the ground truth matrix, both taking values in the interval $[0,1]$. Following the definitions of TP, TN, FP, and FN, the proposed *soft* formulas, operating on continuous values, may be derived by simple min/max operations between the matrices R and G:

$$sTP_{ij} = \min[g_{ij}, r_{ij}] \quad (6)$$
$$sTN_{ij} = \min[1 - g_{ij}, 1 - r_{ij}] \quad (7)$$
$$sFP_{ij} = \max[r_{ij} - g_{ij}, 0] \quad (8)$$
$$sFN_{ij} = \max[g_{ij} - r_{ij}, 0] \quad (9)$$

Similarly to their binary counterparts, the above formulas integrate to unity for every (i, j) element of the two matrices. Fig. 4 demonstrates graphically the new definitions, with bars ranging between 0 and 1, representing the possible values that R and G may acquire. For each definition, two different instances are included, one when the predicted value R is greater than G, and one for the opposite case. The final values of sTP, sTN, sFP, and sFN for the whole database are the sum of all the individual (i, j) cases. The summed values can be used directly in the classic formulas for the calculation of soft Precision (sPrecision), soft Recall (sRecall) and sF1 score, as follows:

$$sPrecision = \frac{sTP}{sTP + sFP} \quad (10)$$
$$sRecall = \frac{sTP}{sTP + sFN} \quad (11)$$
$$sF1 = \frac{2 \cdot sPrecision \cdot sRecall}{sPrecision + sRecall} \quad (12)$$

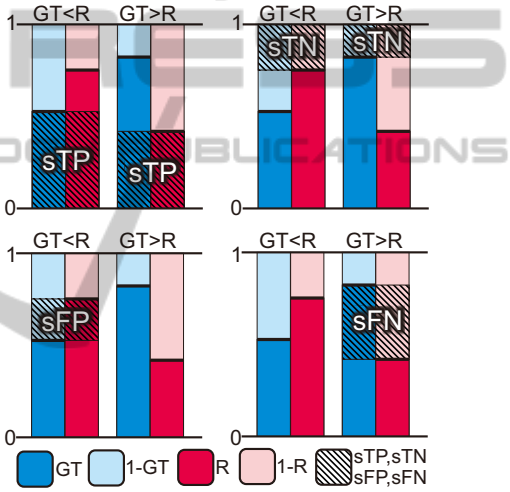

Figure 4: Graphical demonstration of the sTP, sTN, sFP, and sFN definitions.

4 EXPERIMENTAL RESULTS

We compare PhotoCluster to four existing ND detection techniques:

- *Cascade* from Vonikakis and Winkler (2012), a system typical of cascade-based approaches. More importantly, this method is also finely tuned for personal photo collections and uses the same three features as the proposed implementation.
- *INDetector* from Zhang and Chang (2004), which uses stochastic attributed relational graph matching with learning and part-based representation of visual scenes.
- The classic Bag-of-Visual-Words technique (*BoVW*), as implemented by the SOTU ND detection tool (Zhao, 2013). This particular

implementation uses Hessian-Affine keypoint detection and SIFT for keypoint description. The dictionary contains 20,000 visual words and was built using randomly selected keypoints from the same datasets in which the method was tested.

- *Xmatch* from Zhao and Ngo (2009), which uses scale-rotation invariant pattern entropy based on the SIFT descriptor and exhausted pair-wise comparisons.

Apart from these methods, the results for using only the image similarity measure, without the subsequent proposed clustering steps, are also reported (*Similarity*).

Since existing methods mainly target the domain of copyright detection, their evaluation datasets include images from news channels (Zhao and Ngo, 2009), or synthetic degradations (e.g. cropping, rotation, intensity, resizing etc.) applied to images from the web (Foo et al., 2007a). These databases however, are quite different from the personal photolibrary of a typical user, which mostly contains people in family moments, traveling/vacation, or other everyday activities. More importantly though, these datasets provide only a binary ground truth, which cannot capture the ambiguity of NIND cases (Jinda-Apiraksa et al., 2013). For this reason, two different datasets were used for the evaluation of the methods in this study, featuring images taken from personal photolibraries. The comparison results are reported in the following.

4.1 California-ND Dataset

The California-ND dataset has been specifically designed for ND detection in personal photolibraries (Jinda-Apiraksa et al., 2013). The advantage of this dataset is that it comprises 701 images from a real user's travel photo collection, the size of which coincides with the average number of photo taken per trip (Loos et al., 2009). Although the total number of images may not be as high as in other established datasets in the copyright detection domain, this is the *only* existing publicly available dataset including images directly taken from a personal photo collection, which has also been annotated for ND cases by a panel of 10 observers, and as such captures the inherent ambiguity of NIND cases. In order to use them in our evaluation, the 10 annotations were averaged, resulting in a real number in the interval $[0,1]$, indicating the agreement between subjects that a pair of images may be ND. These results are stored in matrix G, which serves as the ground truth.

The ND cases include the 3 major categories reported in Jaimes et al. (2002): variations in the scene, the camera settings, and the image. This includes changes in the subject/background, zooming, panning, tilting, brightness/exposure difference, white balance difference, burst shots, group photos, performance/show photos, portrait photos etc. It should be noted that zooming, in reality, can be different from simple cropping, which is extensively used in other datasets, since by the time the camera lens has zoomed and focused, the scene may also have changed. Furthermore, the photos included in the dataset are captured by two different cameras with non-synchronized timestamps. This has an impact on any method that uses timestamps as a feature of image (dis)similarity, including the proposed one.

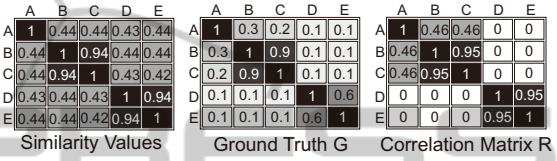

Figure 5: Correlation matrices for the ND cases of Fig. 1.

Fig. 5 demonstrates the strength of PhotoCluster using the images of Fig. 1, which are part of the California-ND dataset. The results of PhotoCluster (matrix R) are compared to the ground truth (matrix G) and the similarity values used in the multiple clustering step. The subset of five images (A, B, C, D, and E), contains one obvious pair of ND (B,C) and many other ambiguous cases. This is evident from matrix G, where the average observers' rating for B and C is 0.9, whereas it ranges from 0.1 to 0.6 for the other pairs.

The image similarity values do not follow the ground truth trend. According to matrix G, only 10% of the observers agreed that images D and E are ND with images A, B, and C, whereas the similarity value for all of them is around 0.44. Once the multiple clustering step of PhotoCluster is applied on these similarity values, R resembles G much better; images D and E are assigned 0 probability of being ND with A, B, and C. This shows that the results of the proposed method roughly follow the pattern of ground truth, whereas image similarity alone is not enough for capturing the ambiguity of NIND cases.

Fig. 6 depicts the performance of the different methods for the California-ND dataset. It confirms again that image similarity alone is not adequate for detecting ND cases; while it exhibits very high recall, it has very low precision. Consequently, the sF1 score is very low. When compared to PhotoCluster, the contribution of the multiple clustering approach becomes apparent.

INDetector and BoVW exhibit very similar results, and a behavior opposite to Similarity. Their pre-

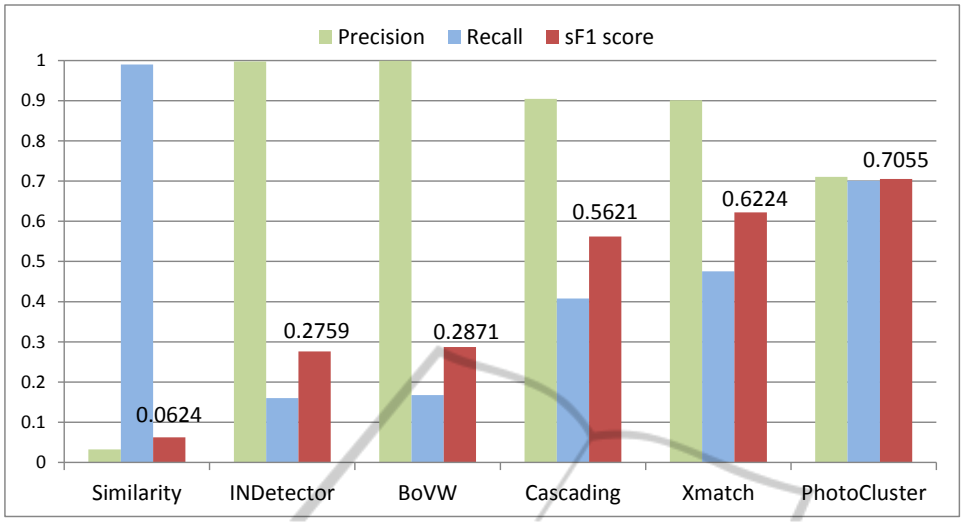

Figure 6: Comparison results for the California-ND dataset.

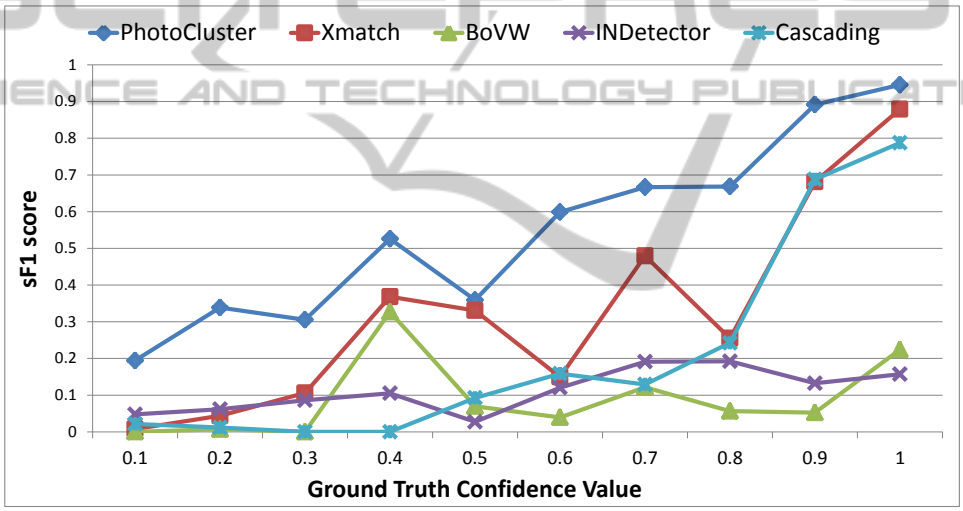

Figure 7: Comparison results for different degrees of ground truth confidence value.

cision is very high (nearly 1), meaning that almost all detected ND cases are correct. However, many ND cases are not detected; this has a profound impact on the recall, which is below 0.2. Consequently, the resulting sF1 score is quite low. This can be explained by the fact that these methods only use local features and do not take other global features like color or time stamps into account, which suggests that local features are not enough for dealing with the ambiguity of NIND cases.

Cascade exhibits an improvement over INDetector and BoVW. There is a drop in precision, but this is compensated by an even greater increase in recall, which raises the sF1 score considerably. This increase in performance could be due to the fact that it is specifically targeted at personal photolibraries (which is not the case for INDetector and BoVW), and because of using additional image features.

Xmatch further improves on the others. Once more, precision drops, but this is compensated by an increase in recall, resulting in a higher sF1 score. Although Xmatch does not use color or timestamp information, its good performance is due to the fact that it employs a sophisticated scale-rotation invariant pattern entropy scheme, which is computationally intensive, however.

PhotoCluster exhibits the highest sF1 score among all methods. Again, the trend is similar; there is a drop in precision combined with an increase in recall, resulting in a higher sF1 score. Increased recall indicates that more ND cases are identified.

Fig. 7 depicts the performance of each algorithm for different degrees of ground truth confidence, revealing important insights regarding their behavior

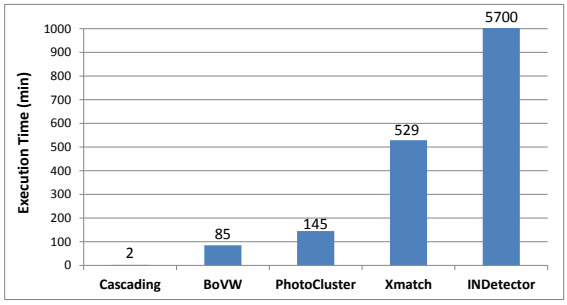

Figure 8: Execution times for the California-ND dataset.

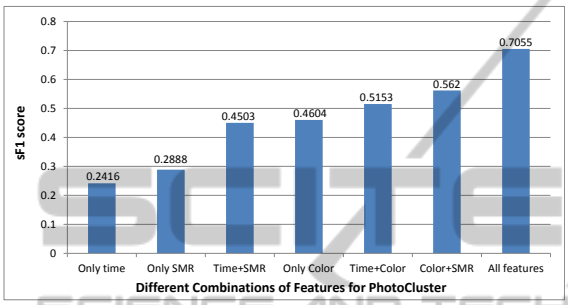

Figure 9: Performance of PhotoCluster for different feature combinations.

for ND cases of different ambiguity level. PhotoCluster exhibits the best sF1 score at all confidence levels, followed by Xmatch. This is a direct result of the fact that other techniques focus on obvious ND pairs, while ignoring uncertain ND cases. For example, at a confidence level of 0.1, PhotoCluster performs 18.65% better in sF1 score compared to Xmatch. This shows that the proposed method can better handle ambiguous ND cases.

The execution time for each method is depicted in Fig. 8. Cascade is the fastest method by far. This is because it quickly eliminates cases with significant differences in color or timestamps, while local feature comparisons are applied only to very few cases. BoVW is the next fastest method. The time for the creation of the dictionary is not taken into consideration, since this is done only once, and any subsequent comparisons reuse it. PhotoCluster is about 3.5 times faster than Xmatch. INDetector is the slowest method by far, to the point where its bar in Fig. 8 is truncated for visualization purposes. Taking the above into consideration, PhotoCluster appears to strike a good balance between high detection performance and moderate execution times. It should be noted that the reported time for PhotoCluster refers to an unoptimized Python implementation, whereas all the other methods were combinations of executable files, built in unknown programming languages.

To further investigate the impact of feature selection on the performance of the proposed method, PhotoCluster was tested using different feature combinations. The results are shown in Fig. 9. Timestamps and SMR are the ones that yield the worst performance. Color histograms, alone or in combination with other features, exhibit better performance compared to SMR and timestamps, while the combination of all three results in the best performance. This indicates that – at least for personal photo collections – local features alone are not enough to successfully identify all ND cases, and a combination of local and global features seems more promising.

4.2 INRIA Holidays Dataset

The proposed method was also tested on INRIA's Holidays dataset (Jegou et al., 2008). Although the target application of this dataset is image retrieval, it was selected due to the lack of other appropriate datasets, as well as the fact that it comprises images taken from personal photo collections. It contains a total of 1491 images and provides 500 queries, along with their ground truth, and a comparison protocol based on mean average precision (mAP), which is given by the following equations:

$$AP = \sum_{k=1}^{n} P(k)\Delta r(k), \tag{13}$$

$$mAP = \frac{\sum_{q=1}^{Q} AP(q)}{Q}, \tag{14}$$

where AP is the Average Precision, k is the rank in the sequence of retrieved images, n is the number of retrieved images (which differs for each query), $P(k)$ is the precision at cut-off k in the list, $\Delta r(k)$ is the change in recall from image $k-1$ to k and Q is the total number of queries, which is 500 for the Holidays dataset.

The results of the different methods on the Holidays dataset are depicted in Fig. 10. PhotoCluster, in its original version, clearly exhibits the best performance. However, the fact that it uses timestamps in this particular dataset skews the results. This is because the images in Holidays dataset are not consecutively chosen from a photolibrary, but instead handpicked from very different time periods. As a result, timestamps become unusually discriminative, giving an unfair advantage to PhotoCluster. For this reason, a second version of PhotoCluster without the use of timestamps was included. It should be noted that this is not a problem for the California-ND dataset, in which all images were selected consecutively from a user's personal photo collection. Furthermore, timestamps were found to be among the less discriminative features in California-ND, according to Fig. 9. This highlights the importance of consecutively selecting

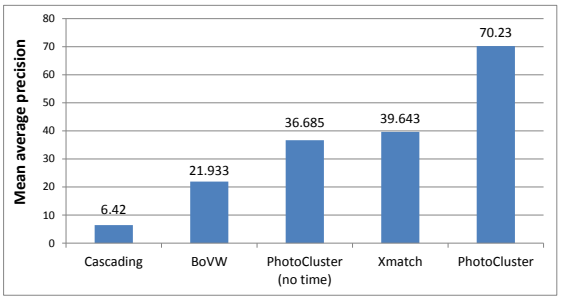

Figure 10: Comparison results for the INRIA Holidays dataset.

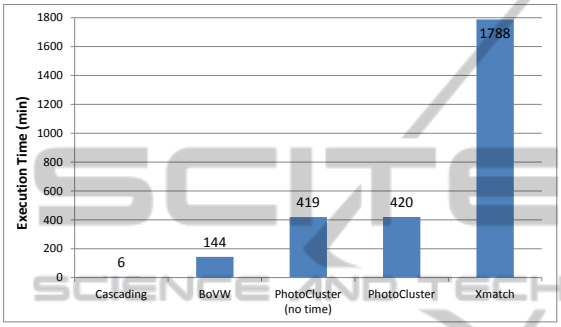

Figure 11: Execution times for the INRIA Holidays dataset.

images from personal photo collections.

The worst performance is exhibited by the Cascading technique, which seems to be affected by the fact that it is finely tuned for the detection of high-confidence ND cases and not for image retrieval. BoVW has a better performance than Cascading. If we exclude the original version of the proposed method, Xmatch exhibits the best results, closely followed by PhotoCluster without timestamps (the difference in mAP is less than 3%). However, PhotoCluster is over 4 times faster than Xmatch, as Fig. 11 shows.

5 CONCLUSIONS

We presented PhotoCluster, a new technique for NIND detection in personal photo collections. The main advantage of the proposed approach is that it produces a correlation matrix indicating the probability that an image pair may be considered ND. As a result, ambiguous ND cases will not be ignored, but will simply be assigned a low probability value. PhotoCluster performs on par or better compared to existing state-of-the-art methods, while keeping complexity and execution times reasonable. Experimental results demonstrate that it can better handle ambiguous cases of ND, which makes it much more suitable for personal photo collections.

ACKNOWLEDGEMENTS

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).

REFERENCES

Chu, W.-T. and Lin, C.-H. (2010). Consumer photo management and browsing facilitated by near-duplicate detection with feature filtering. *Journal of Visual Communication and Image Representation*, 21(3):256–268.

Foo, J. J., Sinha, R., and Zobel, J. (2007a). Discovery of image versions in large collections. In *Proc. 13th International MultiMedia Modelling Conference*.

Foo, J. J., Zobel, J., and Sinha, R. (2007b). Clustering near-duplicate images in large collections. In *Proc. 9th ACM SIGMM International Workshop on Multimedia Information Retrieval*, pages 21–30.

Frey, B. J. and Dueck, D. (2007). Clustering by passing messages between data points. *Science*, 315.

Fu, H., Cao, X., and Tu, Z. (2013). Cluster-based co-saliency detection. *IEEE Transactions on Image Processing*, 22(10):3766–3778.

Jaimes, A., Chang, S., and Loui, A. (2003). Detection of non-identical duplicate consumer photographs. In *Proc. Joint Conference of the 4th International Conference on Information, Communications and Signal Processing, and 4th Pacific Rim Conference on Multimedia*, pages 16–20.

Jaimes, A., Chang, S.-F., and C. Loui, A. (2002). Duplicate detection in consumer photography and news video. In *Proc. 10th ACM International Conference on Multimedia*.

Jegou, H., Douze, M., and Schmid, C. (2008). Hamming embedding and weak geometric consistency for large scale image search. In *Proc. 10th European Conference on Computer Vision (ECCV)*, pages 304–317.

Jinda-Apiraksa, A., Vonikakis, V., and Winkler, S. (2013). California-ND: An annotated dataset for near-duplicate detection in personal photo collections. In *Proc. 5th International Workshop on Quality of Multimedia Experience (QoMEX)*.

Ke, Y., Sukthankar, R., and Huston, L. (2004). Efficient near-duplicate detection and sub-image retrieval. In *Proc. 12th ACM International Conference on Multimedia*, pages 869–876.

Kim, S. J., Ng, H., Winkler, S., Song, P., and Fu, C.-W. (2012). Brush-and-drag: A multi-touch interface for photo triaging. In *Proc. 14th ACM International Conference on Human-Computer Interaction with Mobile Devices and Services (MobileHCI)*, pages 59–68.

Loos, A., Paduschek, R., and Kormann, D. (2009). Evaluation of algorithms for the summarization of photo collections. In *Proc. Theseus/ImageCLEF Workshop on Visual Information Retrieval Evaluation*.

Oliva, A. and Torralba, A. (2001). Modeling the shape of the scene: A holistic representation of the spatial envelope. *International Journal of Computer Vision*, 42(3):145–175.

Schaefer, G. (2010). A next generation browsing environment for large image repositories. *Multimedia Tools and Applications*, 47(1):105–120.

Tang, F. and Gao, Y. (2009). Fast near duplicate detection for personal image collections. In *Proc. 17th ACM International Conference on Multimedia*, pages 701–704.

Vonikakis, V. and Winkler, S. (2012). Emotion-based sequence of family photos. In *Proc. 20th ACM International Conference on Multimedia*, pages 1371–1372.

Wang, M., Ji, D., Tian, Q., and Hua, X.-S. (2012). Intelligent photo clustering with user interaction and distance metric learning. *Pattern Recognition Letters*, 33(4):462–470.

Wang, Y., Hou, Z., and Leman, K. (2011). Keypoint-based near-duplicate images detection using affine invariant feature and color matching. In *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 1209–1212.

Wu, Z., Ke, Q., Isard, M., and Sun, J. (2009). Bundling features for large scale partial-duplicate web image search. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 25–32.

Zhang, D.-Q. and Chang, S.-F. (2004). Detecting image near-duplicate by stochastic attributed relational graph matching with learning. In *Proc. 12th ACM International Conference on Multimedia*, pages 877–884.

Zhao, W.-L. (2013). SOTU. http://www.cs.cityu.edu.hk/ wzhao2/sotu.htm. [Online; accessed 6-Nov-2013].

Zhao, W.-L. and Ngo, C.-W. (2009). Scale-rotation invariant pattern entropy for keypoint-based near-duplicate detection. *IEEE Transactions on Image Processing*, 18(2):412–423.

Zheng, L., Qiu, G., Huang, J., and Fu, H. (2011). Salient covariance for near-duplicate image and video detection. In *Proc. IEEE International Conference on Image Processing (ICIP)*, pages 2537–2540.