ARCADE: Accurate Recognition of Clusters Across Densities

Chad M.S. Steel, Virginia Tech

Abstract

The ARCADE algorithm identifies irregularly shaped clusters of varying density using all possible values of Eps. A novel algorithm based on DBSCAN, ARCADE performs mixed-density clustering similar to k-means without the requirement of known coordinates.

In this paper, the ARCADE algorithm is applied to the problem of pornography detection. Specifically, the ARCADE algorithm is applied to a large dataset comprised of both pornographic and non-pornographic images from multiple websites. The images are clustered using ARCADE and the results of five different colorspaces and two different bin sizes are compared using a sum of one-dimensional Earth Movers Distance histogram measures as a distance function. ARCADE is shown to outperform traditional DBSCAN and produced a 91.58% recall at 80.85% precision using the HSV colorspace.

1. Introduction

As the size of hard drives increase, the need for automated tools to assist in forensic analysis has similarly increased. It is not uncommon for a typical hard drive to contain hundreds of thousands or even millions of images.

One challenge in digital forensic analysis of hard disks is the quick identification of potential pornography (if any) that is present on the drives. To-date, the major digital forensic tool suites do not integrate intelligence into image analysis and simply present a gallery view of all images on a drive that are sortable by meta-data. Alternatively, a hash analysis can be used to identify specific, known images and movies, though the current cryptographic hashsets are vulnerable to small changes in the image. While commercial add-ons such as LTU Finder exist which employ intelligence in their analysis, they require supervised training and a database of known pornography.

Clustering has been successfully used for content-based image retrieval in systems such as CLUE, but these systems have primarily focused on the retrieval of images semantically similar to a reference image and on the subsequent visualization of the resulting subset. *(Smeulders, Worring, Santini, Gupta, & Jain, 2000) (Chen, Wang, & Krovetz, 2005)* The problem of image visualization in digital forensics is related but of a higher computational complexity. Specifically, all of the images present on a drive need to be compared to the other images to be clustered according to series. These images are then grouped for visualization and analysis by a forensic examiner.

By applying a novel density-based clustering technique to an unknown corpus of images, visualization of possible child pornography can be performed in a fraction of the time needed to review the corpus on an image-by-image basis. Individual clusters tend toward coherence (pornographic or non-pornographic), and an examiner can quickly identify groups of suspect images.

ARCADE, an algorithm for the Accurate Recognition of Clusters Across Densities, takes a density-based approach to clustering that is highly applicable to the problem of image-based clustering, as well as the more general problem of clustering where the exact Cartesian coordinates are not known and cannot be easily calculated, but a relative object distance measure is available. ARCADE allows for the identification of clusters of varying densities with noise present using a distance matrix, and the cluster size is user-selectable depending on the application and time available for analysis. ARCADE iterates through all values of Eps using a fixed MinPts value, and identifies clusters in order of density from most-to-least dense.

As applied to the pornography detection, ARCADE is tested using five different colorspaces – RGB, YCbCr, CbCr, CIELAB, and HSV. A simple histogram comparison using a sum of one-dimensional Earth Mover's Distance (EMD) is applied as a distance measure using both 16 and 32 histogram bins per color component. ARCADE is shown to be more effective at clustering than an Eps-optimized DBSCAN, achieving a 91.58% recall of pornographic material with an 80.85% precision using a 16 bin HSV histogram.

2. Prior Art

ARCADE is based on work from two domains – density-based clustering and pornography detection. Density-based clustering is focused primarily on identifying and grouping points based on a single or

small number of fixed densities. Pornography detection focused on identifying objectionable images and has traditionally used color, texture, and shape to identify likely candidate images.

2.1 Density-Based Clustering

Clustering is a mechanism for grouping data into logical partitions. (Kaufman & Rousseeuw, 1990) Many distance based techniques such as k-means and k-medoids are ineffective when absolute coordinates within a reasonably small dimensionality range are not available for the optimization step, as are grid-based techniques. While multi-dimensional scaling allows for the assignment of coordinates, the computation complexity makes it ineffective for large datasets such as may be expected in the image corpus from a hard drive. (Borg & Groenen, 2005)

The original DBSCAN algorithm was designed to detect clusters in a spatial dataset which contained noise. (Ester, Kriegel, Sander, & Xu, 1996) While DBSCAN and its many optimizations are effective at detecting clusters of a similar density, datasets with clusters of multiple densities such as those shown in Figure 1 are difficult to differentiate. Additionally, DBSCAN requires the selection of a single value Eps, indicating the maximum search distance, which requires a priori knowledge of the dataset or substantial pre-processing for maximum effectiveness.

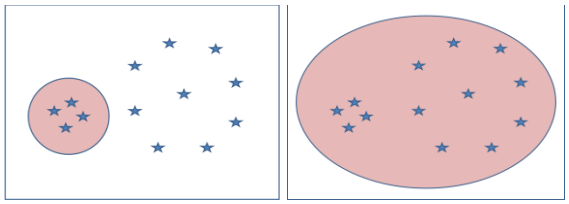

Figure 1. Variable density clusters that cannot be differentiated by DBSCAN.

An optimization to DBSCAN which allow for multiple densities was proposed by Liu et al. under the moniker VDBSCAN. VDBSCAN relies on the visualization of the appropriate jumps in Eps based on a k-distance plot. It was applied successfully to a small, synthetic dataset but required human intervention. Additionally, VDBSCAN was not applied to a large corpus of real. (Liu, Zhou, & Wu, 2007) Pei et al. improved upon the approach used by VDBSCAN with DECODE. Their method identifies thresholds for classification using a reversible jump Markov Chain Monte Carlo (MCMC) strategy. DECODE is algorithmically complex, but was successfully applied to multiple synthetic and actual datasets. DECODE was not designed to be tunable as to the number of clusters, and relies on identifying an overall optimal clustering. (Pei, Jasra, Hand, Zhu, & Zhou, 2009)

2.2 Pornography Detection

A definition of what is pornography must be defined for an accurate comparison of detection techniques. For the purposes of most prior art, the presence of nudity is the criteria for defining an image as pornographic and the same definition is employed in this paper.

One of the earliest detection systems was presented by Forsyth and Fleck, which made use of skin detection in combination with stick-based shape descriptors. Their work showed strong performance for the initial skin classifier, with a 79.3% detection and 11.3% false positive rate. The addition of the geometric classifier resulted in a lower false positive rate of 4.2%, but with a similarly reduced detection rate of 42.7%. *(Forsyth & Fleck, 1999)* A neural network classifier based on simple histogram analysis and skin blobs was implemented by Jones and Rehg. Their classifier was rapid and resulted in an 85.8% detection rate with a 7.5% false positive rate, with additional improvements shown through the use of contextual text. *(Jones & Rehg, 2002)* While the approach in this paper does not utilize contextual text, it could easily be included by modifying the distance measure

For more complex classifiers, a faster histogram approach using a Daubechies' wavelet transforms-based secondary classifier was introduced in WIPE. The classifier was slow, but produced a 96% detection rate on an objectionable dataset with a 9% false positive rate on a benign dataset. The WIPE algorithm relied on template images from a training set to work, however, and was not necessarily broadly applicable to the more general problem of pornographic image detection. *(Wang, Wiederhold, & Firschein, 1997)* In *(Bosson, Cawley, Chan, & Harvey, 2002)*, a k-nearest neighbor classifier was shown to achieve an 87% accuracy, with the added benefit of having a tunable detection approach. In *(Ruiz-del-Solar, Castaneda, Verschae, Baeza-Yates, & Ortiz, 2005)*, a support vector machine classifier is used to achieve a 76% detection rate at an 11% false positive rate on a dataset of approximately 5,000 Internet-obtained images.

3. Dataset

A moderate sized dataset was created from real images present on multiple websites. A web crawler was used to download random files from a large

sample of both pornographic and non-pornographic websites. The resulting images were manually reviewed, and non-pornographic images (due to banners and other images on the websites) were removed from the pornographic images corpus. Additionally, the non-pornographic corpus was confirmed to have no pornographic images present. In total, a corpus of 4,218 images, 2,589 of which were pornographic was created.

The images acquired were in multiple formats, including JPEG, GIF, Bitmap, and PNG. The images ranged in size and complexity from small, black and white icons to multi-megapixel photos. Based on its composition, the corpus is representative of the images expected to be present on the drive of an individual browsing both pornographic and non-pornographic websites.

4. Algorithm

ARCADE can be applied to datasets of varying density without a priori knowledge of the corpus composition. Examples of data compositions that can be detected accurately using ARCADE are shown in Figure 2.

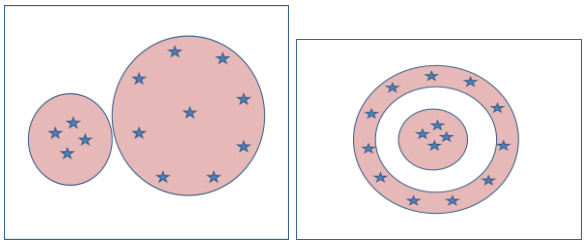

Figure 2. Varying density clusters identifiable by ARCADE

ARCADE was designed to improve upon DBSCAN for datasets with clusters of different densities while retaining the ability of DBSCAN to identify irregularly shaped clusters. The algorithmic design goals which expand upon those used in DBSCAN are as follows:

Different density clusters should be identifiable. ARCADE should not base clustering on a single density estimate that is determined beforehand. Instead, clusters of similar objects which have different densities and shapes should be indentified and differentiated from noise

Knowledge of Cartesian coordinates is not needed. There are multiple applications where the physical coordinates of data points cannot be easily calculated. ARCADE operates based only on the knowledge of distance between data points. The distance function is selectable based on the specific application.

The algorithm is tunable to an approximate number of clusters. The application of image clustering for visualization assumes human review. Based on the time available for review, the approximate number of clusters presented should be configurable. The maximum possible number of clusters should be bounded based on the parameters selected.

Based on these design goals, the ARCADE algorithm for a corpus of items p shown in Figure 3 below. The basic algorithm steps through all possible values of Eps, starting with the smallest, and identifies all clusters within MinPts using DBSCAN. This has the practical impact of identifying clusters of the highest density first (i.e. those within the smallest Eps). As clusters are identified, they are marked and removed from the corpus.

```
Eps_Current=0
Points_Current = p
While Eps_Current < Eps_Max and
Points_Current.Size > Noise_Max
  DBSCAN(Points_Current,Eps_Current,MinPts)
  For i=1 to Points_Current.Size
    If Points_Current(i) != Noise
        Point_Clusters.Add
              (Points_Current(i))
      Points_Current(i).Remove()
    Else
      Points_Current(i) = Unvisited
  Next
  Eps_Current = Eps_Current + Eps_Step
End While
```

Figure 3. ARCADE algorithm for multiple density clustering.

ARCADE takes two parameters, $Noise_{max}$ and MinPts. MinPts is defined as the minimum number of points within a distance Eps that is necessary for a coherent cluster. While MinPts is also used the same as the original DBSCAN, ARCADE additionally uses it to approximate the number of clusters.

The second parameter, $Noise_{max}$, is used as a stopping condition. $Noise_{max}$ is the maximum number of points that may be classified as noise in the corpus. $Noise_{max}$ can be identified empirically by pre-running DBSCAN using the highest tolerable Eps (Eps_{max}), or through user knowledge of the expected number of outliers in the data.

For a corpus of points p:

$$Cluster\_Count_{Max} \leq \frac{|p - Noise_{Max}|}{MinPts}$$

More specifically, Cluster_Count is affected by the granularity in step size for Eps such that:

$$Eps_{Step} \rightarrow 0; Cluster\_Count \rightarrow \frac{|p - Noise_{Max}|}{MinPts}$$

Eps_{Step} is the stepwise increment to Eps. Setting Eps_{Step} too low results in a performance hit, while setting Eps_{Step} too high approximates the original DBSCAN. For the purposes of this paper, Eps_{Step} was set to 1 to cover all possible integer values.

In the unoptimized algorithm, there is a maximum time complexity of $k*p^2$, where k is the maximum number of steps. By pre-generating an ordered list of neighbors at a cost of $p^2/2$, the average complexity can be reduced to k*p*log(p). Additionally, the value of k can be reduced by estimating the lowest possible Eps by ordering the matrix in terms of ascending distance values and finding the smallest run that is less than or equal to MinPts.

Any of the density-based algorithms are highly sensitive to the distance measure used. For the pornography detection application with ARCADE, a simple histogram distance is used. Because many of the images from a given series may have different persons in different positions, shape-based approaches would not be effective. The background and overall color composition is expected to remain the same, however, for images within the same series. Similarly, texture-based and wavelet-based approaches may yield small improvements in matching, but at a complexity cost.

For a histogram matching algorithm the sum of the one-dimensional EMD between histogram component colors is used. EMD calculates the amount of work necessary to transform one histogram into a second. If EMD is applied to the traditional 3-dimensional colorspaces, the computation cost is excessive for this application (Rubner, Tomasi, & Guibas, 2000). However, one-dimensional EMD as applied to each component color is O(n) and can be calculated with the trivial algorithm shown in Figure 4. Calculating the piecewise EMD for each of the three histograms in a given color model and taking the resultant sum yields an effective measure that is quick and tailored to the application.

```
Distance=0
Carry=0
For i=1 to Num_histogram_bins
        Val_a = Hist1[i]
        Val_b = Hist2[i]
        Diff = Val_a - Val_b;
        Distance=Distance+ABS(Diff+Carry)
        Carry=Carry+Diff
Next
```

Figure 4. One-dimensional EMD calculation

In the implementation presented, the distances between images is calculated as needed given the relatively quick speed of EMD calculation and the ability to calculate each of the component colors in parallel. If a large k is expected, the distances can be pre-calculated in an ordered list.

5. Experimental Design

The ARCADE algorithm was applied to the dataset of images collected from the web in an effort to appropriately cluster images into groups that were coherently either pornographic or non-pornographic in composition. Prior to running ARCADE, each of the images on the drive was identified, loaded, and pre-processed.

Image identification for the purposes of the paper consisted of a directory crawl and classification by file extension. Full, forensic implementation would require carving of all deleted content followed by a more complex file signature check to validate the image.

The image pre-processing involved conversion of each of the image formats into a 256-color RGB bitmap. Component RGB histograms were then generated at 16x16x16 and 32x32x32 colors. Individual 3-dimensional histograms for each of the colorspaces evaluated were then created. The individual component histograms were then generated for each of the colorspaces. These component histograms were then normalized to integer values between 0 and 255 to allow for byte-sized storage and reduce overall memory costs.co. For the purposes of these experiments, the histograms were retained, however there is no requirement to retain any of the histogram data following the generation of the distance matrix in an optimized setup.

There were three tests used to evaluate the effectiveness of ARCADE against the image clustering problem. First, a comparison of the relative effectiveness of ARCADE using different colorspaces was performed. Second, a test based on the number of histogram bins was performed. Third, a comparison between ARCADE and traditional DBSCAN using an optimal Eps value was done.

Because pornographic image clustering is largely a skin-based cluster problem, using different colorspaces would be expected to have an impact on overall classification accuracy. Specifically, colorspaces where skin-tone pixel values are more easily separable from non-skin pixel values would be expected to have the best application-specific performance. As such, YCbCr, HSV, and CIELAB would be expected to

perform better at RGB in differentiating skin images (Abadpour & Kasaei, 2005). The experiment uses all of the above colorspaces at a fixed $Noise_{max}$ value of 50 and a fixed MinPts value of 10 (resulting in a relative 10-fold reduction in the number of images to review) to determine the classification accuracy of the various colorspaces.

In addition to the traditional colorspaces, the luminance component in YCbCr is removed and the chrominance values used to determine of the brightness of the image has an impact on classification. The luminance component of skin pixels is not linearly separable as the chrominance components are, and the impact of its inclusion was tested.

The second experiment uses the same colorspaces as the first experiment with two different component histogram sizes – one with 16 bins and one with 32 bins. The bin size directly impacts the comparison speed for the distance measure, therefore a lower bin size is desirable for performance. A larger bin size, however, would be expected to have a greater ability to differentiate histogram differences. Because of the nature of EMD, however, the relative amount of change between histograms would not be expected to be large and it is expected that histogram bin size will have a minimal impact on overall classification accuracy.

The final experiment is a comparison of ARCADE against traditional DBSCAN. For a fixed MinPts (the same value of 10 used in experiments 1 and 2), DBSCAN is run for all possible Eps values using the HSV colorspace and a bin size of 16. Although a priori determination of the optimal Eps cannot be easily determined, the best performing value represents the optimal possible result from the traditional DBSCAN algorithm. The various values of Eps at the fixed MinPts value are then compared for precision, recall, and f-score.

6. Results

The results of the colorspace experiment are shown in Table 1 below. The specific application of pornography detection in the examination of a hard drive generally requires a higher recall than precision – it is better to have the pornographic images retrieved appropriately at the expense of misclassified non-pornographic images. The HSV colorspace appears to have the best recall characteristics (91.58% recall at 80.85% precision) of all of the colorspaces tests at a bin size of 16.

Other applications, such as the filtering of web images, may need a higher precision than recall. The YCbCr colorspace has the highest precision (as well as the highest F-Score), at 84.96%, with a slightly lower recall than HSV of 88.99%. The YCbCr colorspace has the added advantage of a faster calculation time (from RGB) than HSV for speed-intensive applications, though the conversion time is expected to be an order of magnitude lower than the cluster time.

The CbCr colorspace had the lowest overall performance. While the exclusion of luminance may play a role in general skin detection, the consistency in non-skin background pixels present in image series' is shown to be lost when it is removed.

Though HSV had the best performance, the relative difference in performance between colorspaces was minor for this application. Aside from cluster coherence, the closeness of the HSV and CIELAB colorspaces to human visual processing may provide better cluster choice outside of the coherence measurement.

	Precision	Recall	F-Score
CIELAB	80.356%	88.567%	84.262%
RGB	82.934%	89.108%	85.910%
YCbCr	*84.960%*	88.992%	*86.929%*
CbCr	79.128%	89.262%	83.890%
HSV	80.847%	*91.580%*	85.879%

Table 1. Recall and Precision of various colorspaces using ARCADE.

In terms of histogram size, the effect of an increase in the number of bins from 16 to 32 is shown in Figure 5. Relative to precision, HSV and CIELAB showed slight increases as the number of bins increased, while the other colorspaces showed slight decreases. For recall, YCbCr and CbCr showed slight increases (at the expense of precision), while the other colorspaces showed slight decreases. Overall, the effect of bin size was slight on recall and precision performance. It is likely the smaller number of bins more closely simulated a PCA analysis approach, with histogram noise playing a more limited factor. Because of the impact of bin size on performance in terms of both memory usage and speed, a bin size of 16 is recommended based on the results.

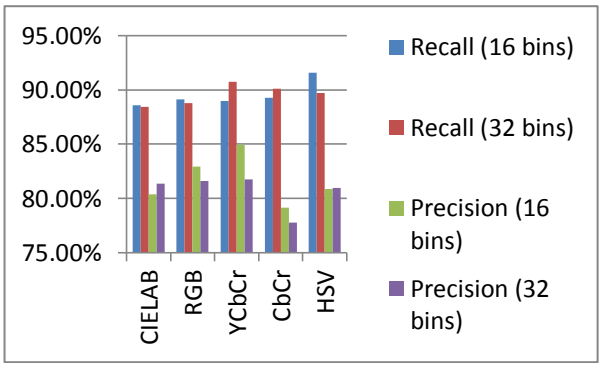

Figure 5. Differences in precision and recall at 16 and 32 bins using ARCADE.

A comparison of the results of ARCADE to traditional DBSCAN using the HSV colorspace and 16 bins required identifying the optimal Eps value for the algorithm. All possible Eps values were iterated through, and the highest F-Score identified. A high F-Score value of 81.62% (80.99% Precision, 82.27% Recall) was identified at an Eps value of 737 as shown in Figure 6. This is compared to an F-Score of 85.88% achieved with ARCADE using the same parameters. ARCADE outperformed DBSCAN by a significant margin in precision, recall, and F-Score at an optimal Eps. Given the complexity in identifying the optimal Eps value, ARCADE is a further improvement over DBSCAN in overall speed at the optimal value.

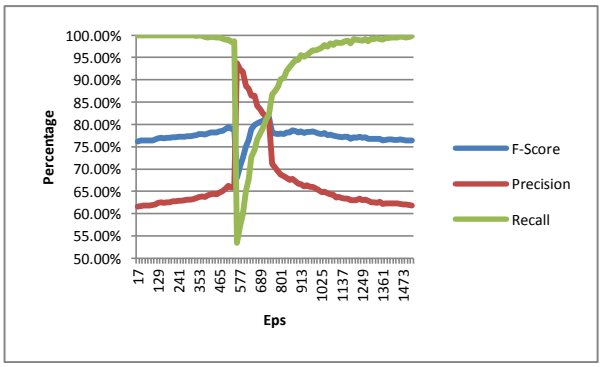

Figure 6. Performance of DBSCAN over different Eps values.

As a side benefit for manual review, the clustering tended to place individual images from the same series together within a cluster. On a real drive with greater coherence between images, the series groupings may provide pointers to specific caches of images.

Examples of clusters identified by ARCADE with the HSV colorspace at a bin size of 16 are shown in Appendix A. Examples of coherent non-pornographic and pornographic image clusters are shown the first two series, respectively. The first cluster returned primarily landscapes with muted blue tones. The second cluster returned pornographic images of the same two individuals within the same series. The final cluster is predominantly non-pornographic, but a misidentified pornographic images is included due to the overwhelming intensity of the background and similarity to the predominant foreground coloring in the other images.

7. Conclusions

The clustering of images as applied to the problem of digital forensics presents a unique set of challenges. There are traditionally a large number of images whose composition is not known a priori, and human review is generally used for confirmation.

ARCADE is shown to be effective in clustering images and specifically the differential grouping of pornographic images. Pre-processing can be performed on the corpus by using a high recall skin detection algorithm to remove obvious no-skin images. The groupings in ARCADE can additionally be coupled with a traditional pornography detection algorithm to rank the most likely pornographic images in each cluster. This allows a forensic examiner the ability to review a much smaller number of images, specifically the most likely for each cluster.

The ability to select an approximate number of clusters is critical in the case of digital forensics. When searching a hard drive for the presence of child pornography, a customs examiner may have a few minutes and need to focus on the most likely candidates from a very small number of clusters. On the other side, an examiner in a lab setting may have several hours to review a drive in detail. ARCADE allows for both scenarios by tuning to the expected number of clusters.

For traditional image grouping, ARCADE was shown to outperform DBSCAN on a real dataset. The algorithm can be applied to any other density-based clustering problem, and only requires a relative distance measure – Cartesian coordinates are not necessary. On a single drive with images that are related such as vacation photos, downloads from the same website, or movie stills, the performance of ARCADE would be expected to be even higher than that on the sample set.

8. References

Abadpour, A., & Kasaei, S. (2005). Pixel-Based Skin Detection for Pornography Filtering. *Iranian Journal of Electrical and Electronic Engineeering , 1* (3).

Borg, I., & Groenen, P. (2005). *Modern Multidimensional Scaling: Theory and Applications.* Springer.

Bosson, A., Cawley, G. C., Chan, Y., & Harvey, R. (2002). Non-retrieval: Blocking Pornographic Images., (pp. 50-60).

Chen, Y., Wang, J. Z., & Krovetz, R. (2005). CLUE: cluster-based retrieval of images by unsupervised learning. *Image Processing, IEEE Transactions on , 14*, 1187-1201.

Ester, M., Kriegel, H.-P., Sander, J., & Xu, X. (1996). A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise., (pp. 226-231).

Forsyth, D. A., & Fleck, M. M. (1999). Automatic detection of human nudes. *International Journal of Computer Vision , 32*, 63--77.

Jones, M. J., & Rehg, J. M. (2002). Statistical Color Models with Application to Skin Detection. *International Journal of Computer Vision , 46*, 81-96.

Kaufman, L., & Rousseeuw, P. J. (1990). *Finding Groups in Data: An Introduction to Cluster Analysis.* John Wiley & Sons.

Liu, P., Zhou, D., & Wu, N. (2007). VDBSCAN: Varied Density Based Spatial Clustering of Applications with Noise. *Service Systems and Service Management, 2007 International Conference on*, (pp. 1-4).

LTU Technologies. (2009, June). Retrieved June 15, 2009, from LTU Imageseeker: http://www.ltutech.com/en/products/ltu-finder

Pei, T., Jasra, A., Hand, D. J., Zhu, A., & Zhou, C. (2009). DECODE: a new method for discovering clusters of different densities in spatial data. *Data Min. Knowl. Discov. , 18*, 337--369.

Rubner, Y., Tomasi, C., & Guibas, L. J. (2000). The Earth Movers Distance as a Metric for Image Retrieval. *Int. J. Comput. Vision , 40*, 99--121.

Ruiz-del-Solar, J., Castaneda, V., Verschae, R., Baeza-Yates, R., & Ortiz, F. (2005). Characterizing objectionable image content (pornography and nude images) of specific Web segments: Chile as a case study. *Web Congress, 2005. LA-WEB 2005. Third Latin American*, (pp. 10 pp.-).

Smeulders, A. W., Worring, M., Santini, S., Gupta, A., & Jain, R. (2000). Content-based image retrieval at the end of the early years. *Pattern Analysis and Machine Intelligence, IEEE Transactions on , 22*, 1349-1380.

Wang, J. Z., Wiederhold, G., & Firschein, O. (1997). System for Screening Objectionable Images Using Daubechies Wavelets and Color Histograms. (pp. 20--30). Springer-Verlag.

Appendix A. Sample Image Clusters

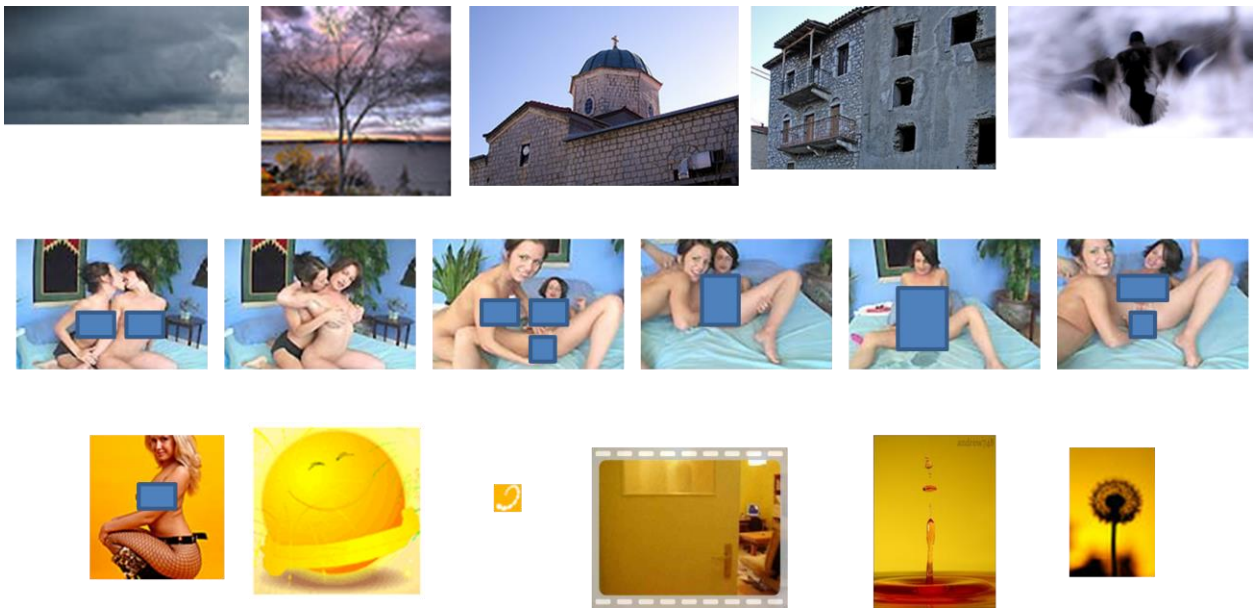

Figure 8. Examples of a cluster of non-pornographic images, a cluster of pornographic images, and a non-pornographic cluster with a single misclustered images (boxes added).