Enhancement, Preprocessing, and Machine Learning with Galaxy Images

JOHN JENKINSON, ARTYOM GRIGORYAN, SOS AGAIAN

SPIE 2015 CONFERENCE ON ELECTRONIC IMAGING

Electrical Engineering, University of Texas at San Antonio

Overview

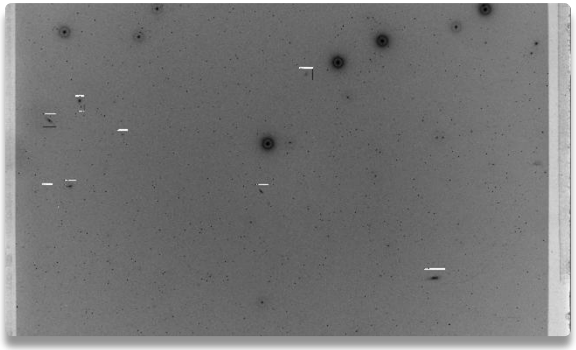
- Data Type & Collection
- Problem Statement & Motivation for Work
- Heap Transform Enhancement
- Image Preprocessing
- Hubble Classification Scheme
- Feature Extraction
- Support Vector Machine Learning
- Principal Component Analysis
- Classification Results
- Conclusion

Data Type & Collection

- Optical galaxy images belong to the Tonantzintla Digital Sky Survey, which is a catalog of images taken by the Camera Schmidt, Figure 1., starting its operation in 1942.
- The spherical mirror of the Camera Schmidt is 762 mm in diameter and coupled to a 660.4 mm correcting plate. The 8x8 inch^2 photographic plates cover a 5°x5° field with a plate-scale of 95 arcsec/mm.
- The plates are first digitized at the maximum optical resolution of the scanner, 4800 dots per inch (dpi), and then rebinned by a factor 3 for a final pixel size of ~ 15 μm (1.51 arcsec/pixel) and transformed to the transparency (positive) mode. Each image has 12470 x 12470 pixels (about 350 Mb in 16-bit mode) and is stored in FITS format.

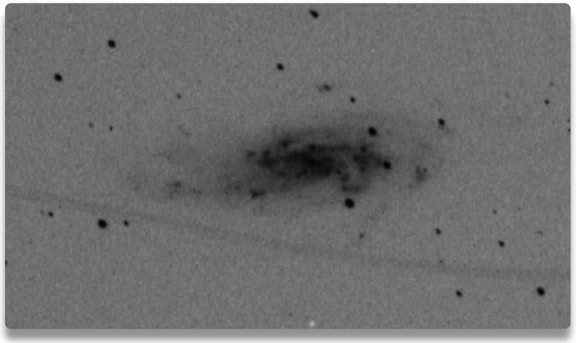
Figure 1. Camera Schmidt

Data Type & Collection



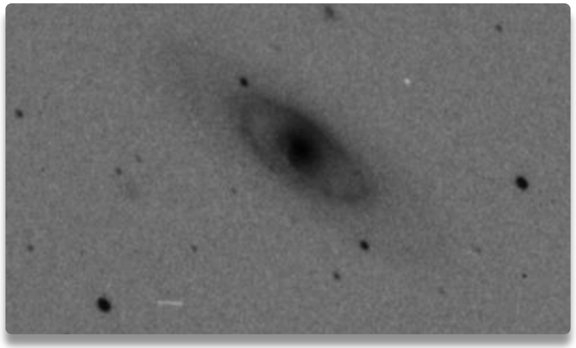
AC 8409 Marked

Digitized plan scans were provided by the Institute of Astrophysics, Optics, and Electronics, in Tonantzintla, Puebla, Mexico, with all galaxies in the image marked and labeled.



NGC 4559 Extracted

Processing entire plate scans by algorithms such as the Watershed for segmentation resulted in memory exhaustion. Therefore, each galaxy was extracted individually for further processing.



NGC 4274 Extracted

NGC 4559 and NGC 4274 are examples of galaxies that have been extracted from the digital plate scan AC 8409.

Problem Statement & Motivation

- Many galaxies contain faint features, such as the spiral arms of NGC 4258 in Figure 2. These faint features are destroyed during segmentation, thereby increasing classification error.
- Faint features are either closely resembling background intensities or are sparse in density.
- Enhancement poses a solution for emphasizing the faint features by differentiating them from background intensities, thereby preserving a more accurate representation of the galaxy post segmentation and decreasing classification error.

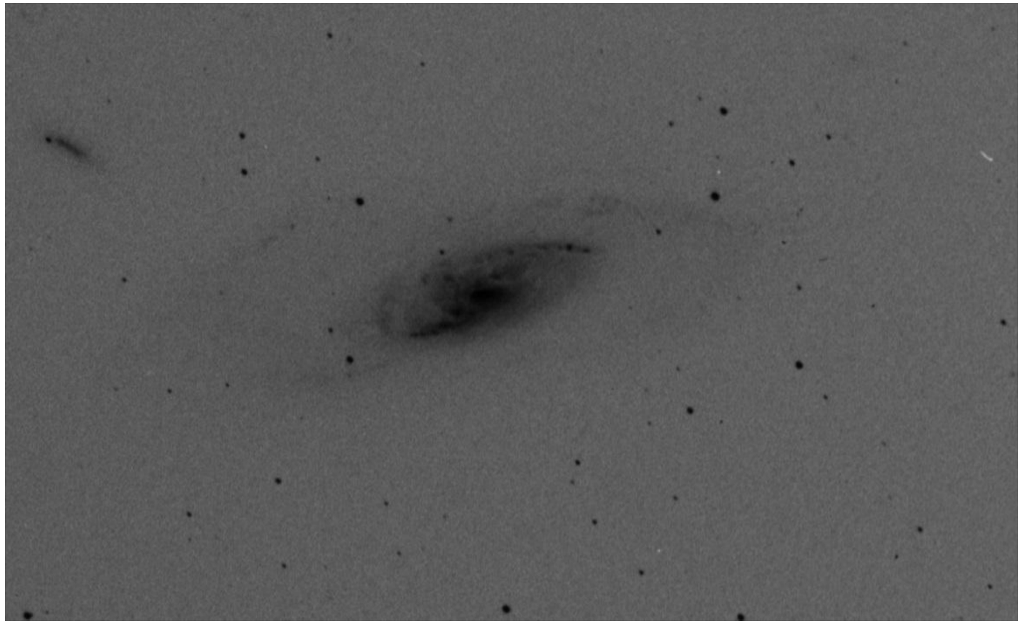


Figure 2. NGC 4258 appears to have faint spiral arms.

Heap Transform Enhancement

- Unlike the Fourier, cosine, Haar, etc… transforms which have defined basis functions, the Heap transformation requires the specification of a "signal generator" for processing.
- The Heap transform is defined by the system of decision equations (1), which are used to find different angles φ (2) that define the heap transform at each stage of the transformation. At each stage, the values of x are selected in some order from the signal generator, and produce angle φ and y_0 for a user specified value of a. The angle φ is then used to generate the Heap transform at the first stage, T_{φ_1}, which is defined by Given's rotation matrix (3). T_{φ_1} is then applied to the first two points of the input signal z (4). The superscripts of the output indicate the number of times a specific point has been processed.

$$\begin{aligned} f(x,y,\varphi) &= x_0\cos(\varphi) - x_1\,sin(\varphi) = y_0 \\ g(x,y,\varphi) &= x_0\,sin(\varphi) - x_1\cos(\varphi) = a \end{aligned} \qquad (1)$$

$$\varphi_1 = arccos\left(\frac{a_1}{\sqrt{x_0^2 + x_1^2}}\right) \qquad (2)$$

$$T_{\varphi_1} = \begin{bmatrix} \cos(\varphi_1) & -sin(\varphi_1) \\ sin(\varphi_1) & \cos(\varphi_1) \end{bmatrix} \qquad (3)$$

$$[T_{\varphi_1}]\begin{bmatrix} z_0 \\ z_1 \end{bmatrix} = \begin{bmatrix} z_0^{(1)} \\ z_1^{(1)} \end{bmatrix} \qquad (4)$$

Heap Transform Enhancement

- The process of generating rotation transforms from the signal generator is repeated until all of the points of the input signal have been processed. This stage-wise transform is illustrated in Figure 3.
- For galaxy image enhancement, the median of each row in the image was selected to be the signal generator for that row of the image. Each row of the image was then processed as a 1-Dimensional signal.

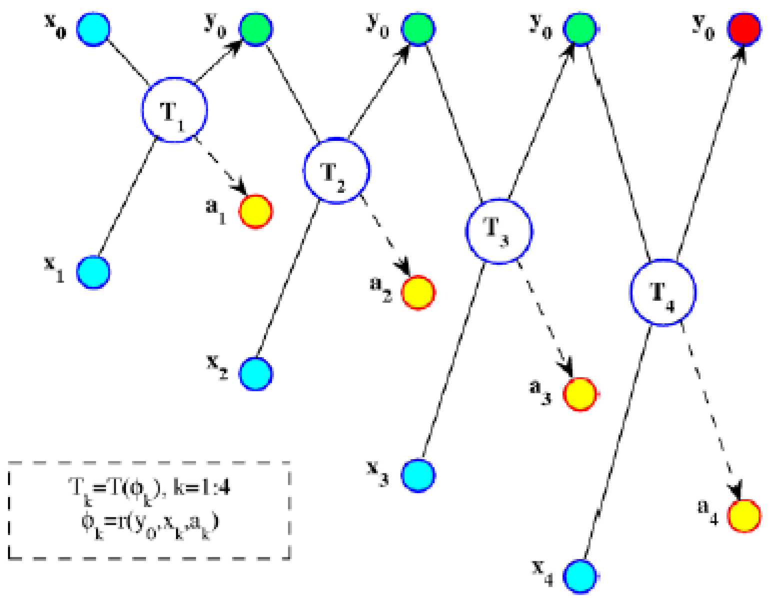


Figure 3. Signal-flow graph of determination of the five-point transform by a vector $x=(x_0,x_1,x_2,x_3,x_4)'$.

Image Preprocessing

- For the image *f(x,y)*, the following operations were applied.
- Thresholding $g(x,y) = \begin{cases} 1, if\ f(x,y) > T \\ 0, otherwise \end{cases}$ for some value of T.
- Opening $O(x,y) = f(x,y) \circ B = \bigcup\{B + z : B + z \subset f(x,y)\}$ where B is the disc structuring element and z is a point in the image *f*.
- Rotation by the angle defined by image second moments. $\theta = \left(\frac{2\mu_{11}}{\mu_{20} - \mu_{02}}\right)$
- Shifting by the vector from the galaxy centroid to the image center. $\left(\frac{\sum_n \sum_m n f_{n,m}}{\sum_n \sum_m f_{n,m}}, \frac{\sum_n \sum_m m f_{n,m}}{\sum_n \sum_m f_{n,m}}\right)$

Image Preprocessing

- All images were resized to a uniform 128x128 pixels.
- Canny edge detection was used to detect galaxy edges.
- Bounding Box and Best Fit Ellipse were calculated for each galaxy.
- Figure 4. shows the original image and all processing steps.

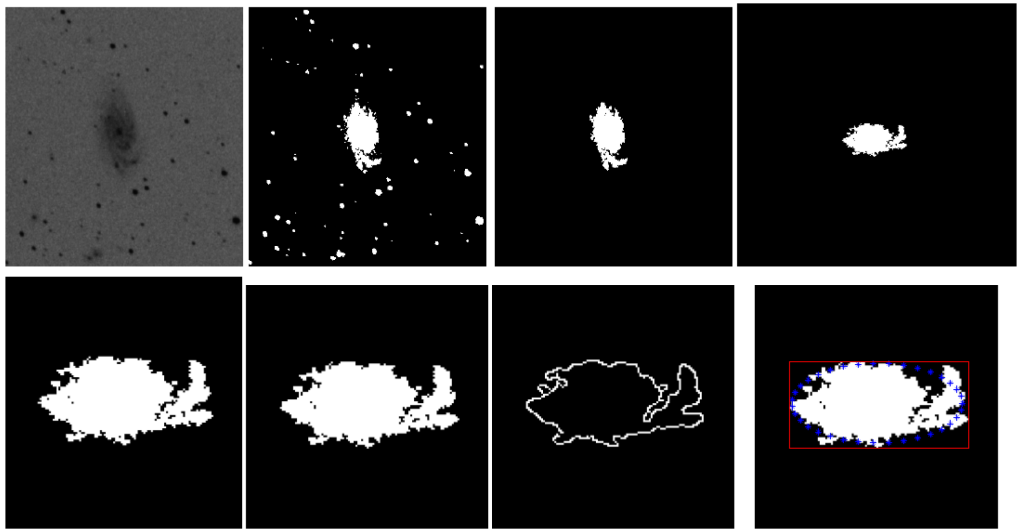


Figure 4. Original image, thresholding, opening, rotation, centering, resizing, edge detection, bounding box and best fit ellipse.

Hubble Classification Scheme

- Galaxy images were classified into the classes Elliptical (E), Lenticular (S0), Spiral (S), Barred Spiral (SB), and Irregular (Irr).
- Classification was performed class-pair wise so that first galaxies were classified as Irregular or Regular, then Irregular galaxies were removed from the training and test set. Next, galaxies were classified as Elliptical or not Elliptical, and so on.

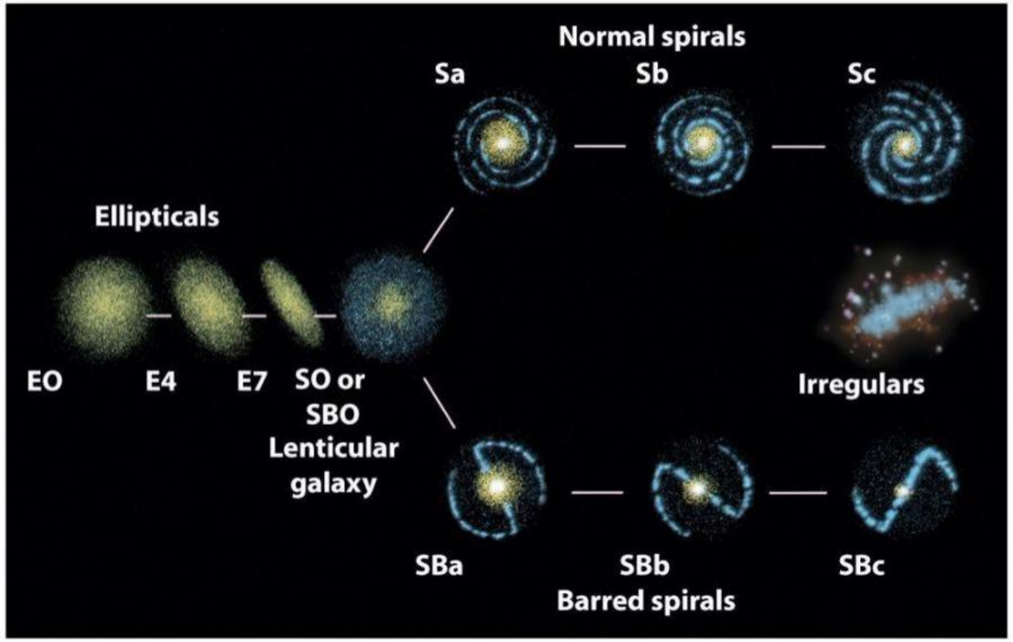


Figure 5. Hubble Classification Scheme

Feature Extraction

Elongation | Form Factor | Convexity | Bounding Box to Fill Factor | Bounding Box to Perimeter | Asymmetry Index

Feature	E	F	C	BFF	BP	AI
Formula	$(a-b)/(a+b)$	A/P^2	$P/(2H+2W)$	A/HW	$HW/(2H+2W)^2$	$\sum_{i,j}\lvert I(i,j)-I_{180}(i,j)\rvert/\sum_{i,j}\lvert I(i,j)\rvert$

Features	Elliptical	Lenticular	Simple Spiral	Barred Spiral	Irregular
E	0.071	0.382	0.547	0.485	0.214
F	0.059	0.049	0.025	0.029	0.044
C	0.888	0.872	1.05	1.01	0.953
BFF	0.744	0.699	0.609	0.583	0.634
BP	0.062	0.052	0.043	0.048	0.059
AI	0.274	0.375	0.510	0.464	0.354

Features	Elliptical	Lenticular	Simple Spiral	Barred Spiral	Irregular
E	0.061	0.3914	0.522	0.451	0.199
F	0.041	0.045	0.029	0.030	0.031
C	1.06	0.886	1.00	1.03	1.18
BFF	0.689	0.666	0.581	0.600	0.630
BP	0.062	0.052	0.045	0.050	0.060
AI	0.394	0.290	0.463	0.474	0.659

Support Vector Machines

- Class discrimination by $f(\boldsymbol{x}) = \boldsymbol{w}^T \boldsymbol{x} + b$ and decision boundary of $\{\boldsymbol{x}: f(\boldsymbol{x}) = 0\}$.
- The margin can be written as $M_D(f) = \frac{1}{2}||\mathbf{w}||[\boldsymbol{w}^T \boldsymbol{x}_+ - \boldsymbol{w}^T \boldsymbol{x}_-] = \frac{1}{||\mathbf{w}||}$, where $\mathbf{w}$ is a unit vector.
- Non linearly separable data is mapped to a feature space where it is linearly separable by a kernel function $\phi: \Re \longrightarrow \Im$, then $f(\boldsymbol{x}) = \phi^T(\boldsymbol{x})\phi(\boldsymbol{x}) + b$.
- Kernels used were linear $d = 1$ and

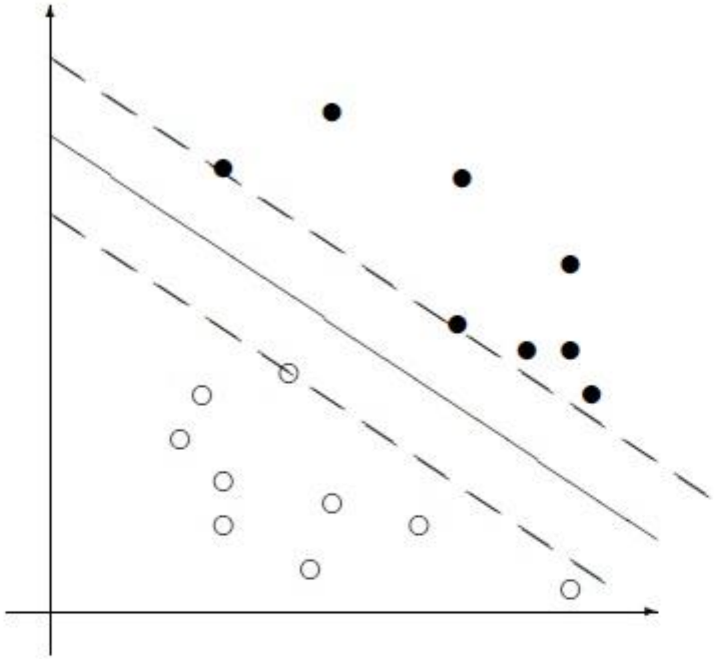


Figure 6. Linearly separable data with decision boundary and maximum margin.

Principal Component Analysis

- Principal Component Analysis (PCA) transforms the original data into equivalent uncorrelated data so that the covariance matrix of the new data is diagonal and the diagonal entries decrease from top to bottom.
- For the data set x_i, with N observations and K features written as the $N \times K$ matrix X, the covariance matrix is $C_X = \left(\frac{1}{N-1}\right) X^T X$. PCA finds R such that, $Y = XR$ and $C_Y = R^T X^T XR = R^T C_X R$. The first column r_1 of R is the first principal component, and can be derived using Lagrangian multipliers, $\phi(r_1, \lambda) = r_1^T C_X r_1 - \lambda_1 (r_1^T r_1 - 1)$. With $\frac{\delta\phi(r_i,\lambda)}{\delta r_i}$ set equal to zero, $C_X r_1 - \lambda_1 r_1 = 0$. This shows that $\lambda_1$is an eigenvalue of C_X and equates to maximizing the variance along the first principal component. The remaining principal components are derived in the same manner.
- Figure 7. shows the classification of classes Irregular vs. Regular with the original 6-

Principal Component Analysis

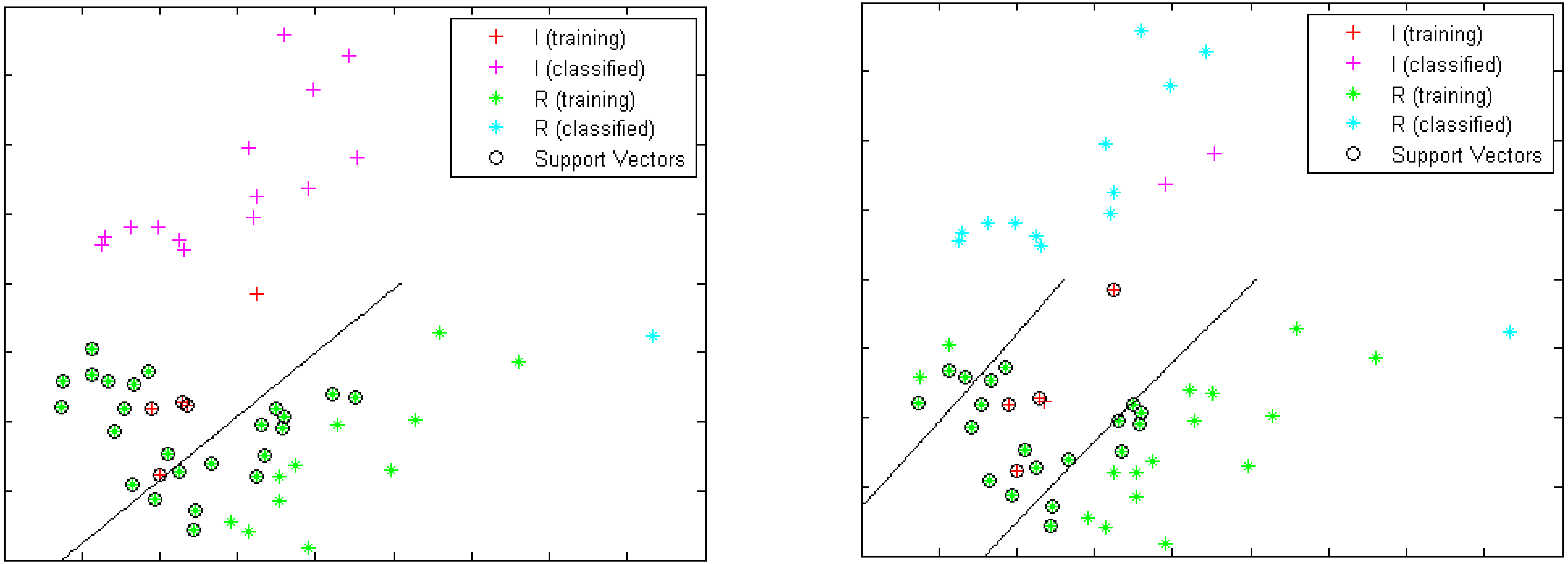


Figure 7. Irr/Reg classification in PCA feature space using left: linear kernel and right: quadratic kernel.

Results

	# test images	# correctly classified (%)			
		6 features		2 PCA features	
		Original	Enhanced	Original	Enhanced
Linear kernel					
Irregular/Regular	15	7 (46.7%)	4 (26.7%)	2 (13.3%)	2 (13.3%)
Elliptical/Not Elliptical	15	13 (86.7%)	11 (73.3%)	3 (20.0%)	10 (66.7%)
Lenticular/Spiral	13	11 (84.6%)	11 (84.6%)	9 (69.2%)	9 (69.2%)
Spiral/Barred Spiral	9	7 (77.8%)	8 (88.9%)	2 (22.2%)	7 (77.8%)
Quadratic kernel					
Irregular/Regular	15	13 (86.7%)	12 (80.0%)	12 (80.0%)	0 (0.0%)
Elliptical/Not Elliptical	15	10 (66.7%)	12 (80.0%)	3 (20.0%)	13 (86.7%)
Lenticular/Spiral	13	8 (61.5%)	11 (84.6%)	3 (23.1%)	9 (69.2%)
Spiral/Barred Spiral	9	4 (44.4%)	6 (66.7%)	2 (22.2%)	6 (66.7%)

Table 5. Summary of classification results for original and enhanced data. Enhancement increased the accuracy by 13.1%.

Conclusion

- Enhancement of galaxy images improved the overall performance of classification.
- Locally, enhancement can degrade performance of classification. This is likely due to intensity variations between original and enhanced images causing segmentation error at the thresholding stage of preprocessing, since the same threshold values were used for both data sets.
- The quadratic kernel in SVM and PCA both improve classification of galaxies for some pairs.
- Further investigation is needed to determine best threshold selection for data after enhancement, and for which pair-wise classifications performance is highest for linear/quadratic kernel and PCA or original feature space.

References

- Hubble, E. P., "Extragalactic nebulae.," Astrophysical Journal **64**, 321–369 (Dec. 1926).
- Storrie-Lombardi, M. C., Lahav, O., Sodre, Jr., L., and Storrie-Lombardi, L. J., "Morphological Classification of Galaxies by Artificial Neural Networks," Monthly Notices of the Royal Astronomical Society **259**, 8P (Nov. 1992).
- Raquel Díaz-Hernández; J. Jesús González; Rafael Costero; José Guichard, "*Retrieval of spectroscopic information from the Tonantzintla Schmidt camera archival plates*," Proc. SPIE 8011, 22nd Congress of the International Commission for Optics: Light for the Development of the World, 80117Z (3 November 2011); doi: 10.1117/12.903386.
- Grigoryan, A.M. (2014) New Method of Givens Rotations for Triangularization of Square Matrices. *Advances in Linear Algebra & Matrix Theory*, **4**, 65-78. http://dx.doi.org/10.4236/alamt.2014.42004.
- Grigoryan, A. M. and Hajinoroozi, M., "A novel method of filtration by the discrete heap transforms," (2014).
- Edward R. Dougherty and Jaakko T.Astola. An Introduction to Nonlinear Image Processing. TT16 SPIE Press (1994).
- Zeljko Ivezic, Andrew J. Connolly, Jacob T. VanderPlas, and Alexander Gray. Statistics, Data Mining, and Machine Learning in Astronomy. Princeton University Press (2014).