Computer-generated visual morphology catalog of ∼3,000,000 SDSS galaxies

Evan Kuminski and Lior Shamir*

Department of Computer Science, Lawrence Technological University, MI 48075
*email: lshamir@mtu.edu

ABSTRACT

We applied computer analysis to classify the broad morphological type of $\sim 3 \cdot 10^6$ SDSS galaxies. The catalog provides for each galaxy the DR8 object ID, right ascension, declination, and the certainty of the automatic classification to spiral or elliptical. The certainty of the classification allows controlling the accuracy of a subset of galaxies by sacrificing some of the least certain classifications. The accuracy of the catalog was tested using galaxies that were classified by the manually annotated *Galaxy Zoo* catalog. The results show that the catalog contains ∼900,000 spiral galaxies and ∼600,000 elliptical galaxies with classification certainty that has a statistical agreement rate of ∼98% with *Galaxy Zoo* debiased "superclean" dataset. The catalog also shows that in rare cases objects assigned by SDSS pipeline with a relatively high redshift (z>0.4) can have clear visual spiral morphology. The catalog can be downloaded at `http://vfacstaff.ltu.edu/lshamir/data/morph_catalog`. The image analysis software that was used to create the catalog is also publicly available.

Subject headings: catalogs — techniques: image processing — methods: data analysis

1. Introduction

Autonomous digital sky surveys have been becoming increasingly dominant in astronomy research, and the trend is bound to continue (Borne 2013; Djorgovski et al. 2013). Digital sky surveys have powerful image acquisition capabilities, allowing them to collect and store data of very large image databases of astronomical objects. One of the bottlenecks in processing astronomical data is the analysis of the morphology of celestial objects. While substantial information about an astronomical object lays within its morphology and visual structure, automatic morphological analysis requires advanced computer vision algorithms, making the analysis of millions of astronomical objects a challenge.

Several methods for automatic galaxy image analysis have been proposed, includ-

ing GALFIT (Peng et al. 2002), GIM2D (Simard 1999), CAS (Conselice 2003), Gini (Abraham et al. 2003), Ganalyzer (Shamir 2011), and SpArcFiRe (Davis and Hayes 2014), as well as image analysis methods based on machine learning (Shamir 2009; Huertas-Company et al. 2009; Banerji et al. 2010; Kuminski et al. 2014; Dieleman et al. 2015).

One of the goals of these methods is to analyze the morphology of galaxies in large galaxy image databases, so that the morphological information can be part of data releases and used similarly to the other information collected by digital sky surveys. However, artifacts, imperfect images, limited resolution, saturated objects, small object surface size, and misidentified objects introduce a challenge when applying these methods to produce clean automatically generated catalogs of galaxy morphology.

Manual inspection of galaxy images can provide accurate annotation, but due to the labor-intensive nature of the task such catalogs can practically contain several thousand galaxies (Baillard et al. 2011; Nair and Abraham 2010). To allow the analysis of larger databases, manual annotation was carried out by non-astronomer volunteers, accessing the images and submitting their annotation via the internet (Lintott et al. 2008) to generate catalogs of several hundred thousands annotated galaxies such as Galaxy Zoo 1 (Lintott et al. 2011) and Galaxy Zoo 2 (Willett et al. 2013). However, even when using the power of citizen scientists the annotation process is still limited by the amounts of data it can analyze, requiring several years and substantial labor to produce cat-

alogs such as Galaxy Zoo. That bandwidth problem will be magnified when more powerful sky surveys such as the Large Synoptic Sky Survey (Abell et al. 2009) see first light, and acquire billions of galaxy images.

Using non-professional astronomers as data annotators, without the ability to control their skill, experience, and quality of their individual annotations, clearly adds noise into the system. Although the presence of malicious annotators is mild (Lintott et al. 2011), participants who are not trained scientists might make wrong or careless annotations, making the data unclean. To correct for human error it is required to apply statistical methods that ignore objects on which the voting was not consistent across the data annotators. In Galaxy Zoo, a classification of a galaxy is defined "clean" if 80% or more of the voters who classified that galaxy provided the same answer (Lintott et al. 2011). A classification of a galaxy is defined "superclean" if it has an agreement of 95% or more of the votes. Filtering classifications that did not meet a certain level of agreement threshold can provide a cleaner subset of the data, but at the cost of sacrificing some of the samples that their voting did not meet the "clean" or "superclean" criteria. For instance, for question 1 (''Is the galaxy simply smooth and rounded, with no sign of a disk?") in Galaxy Zoo 2, just $\sim$7.8% of the galaxies that were classified in the entire dataset met the "superclean" criterion (Kuminski et al. 2014). That is, an average number of 44 citizen scientists (Willett et al. 2013) voted on each galaxy to indicate whether it is smooth and rounded or not, but different citizen scientists provided different answers on the

same question and the same galaxy. Only in ∼7.8% of the galaxies that were classified the agreement among the voters was 95% or higher. That shows that even when using a very high number of data annotators not all celestial objects can be classified with conclusive results.

The inability of manual annotation to perform exhaustive analysis of extremely large databases reinforces the need for automatically generated galaxy morphology catalogs. Simard et al. (2011) performed and released data from automatic analysis of Sersic and disk models of over 10^6 SDSS galaxies. Huertas-Company et al. (2010) released an automatically generated catalog of Sloan Digital Sky Survey galaxies with spectra, and classified them into four basic Hubble morphological types of E, S0, Sab, and Scd. That work was followed by an automatically generated catalog of visual morphologies of ∼50,000 CANDELS galaxies (Huertas-Company et al. 2015a,b).

Here we applied computer vision and pattern recognition methodology to analyze the broad morphology of ∼3 million galaxies taken from Sloan Digital Sky Survey (SDSS) DR8. The catalog is publicly available, as well as the source code of the method that was used to analyze the galaxy images. The methodology can be used to analyze the broad morphology of existing and future larger databases of galaxy images such as the Large Synoptic Survey Telescope.

2. Methods

The initial set of galaxies was selected from SDSS (York et al. 2000) DR8 through the Catalog Archive Server (CAS) based on the constraints described in Table 1.

The query aimed at selecting just galaxies of a sufficiently large size that allows the analysis of their morphology, and avoided as much as possible saturated images or images with multiple peaks or Petrosian radii that could be galaxy mergers or multiple sources that make the morphology of the object more complex.

The vast majority of the objects returned by the query did not have spectra. However, many of the objects with spectra are small or faint, and therefore often do not allow good visual identification of the morphology. Larger objects are likely to lead to a higher identification rate, as larger objects contain more noticeable visual cues that can be used for its correct morphological classification.

The query returned 2,939,891 objects, and the SDSS image of each object was downloaded automatically through CAS. The JPG images were saved in the color TIF (Tagged Image File) format, and each image had the dimensionality of 125×125 pixels. To avoid pressure on CAS server, only one image was downloaded at a time. Downloading the images of all objects from CAS was completed after ∼25 days.

To handle some of the larger objects that their size exceeded the size of the image, for each image that was downloaded the pixels on the edge of the image that had grayscale value greater than 125 were counted. The initial scale of the image is 0.1 arcseconds per pixel. If more than 25% of the edge pixels have grayscale value greater than 125 it is assumed that the edges are not background sky, and therefore the object is too large to fit inside the image. In that case the scale was changed

3

Table 1: Attributes of the CAS query used to select the initial set of galaxies

Field	Constraint	Description
type	=3	Objects identified as galaxies by SDSS pipeline
petroRad_r	>5.5	Petrosian radius greater than 5.5 arcseconds
petroRadErr_r	<5	Petrosian radius error smaller than 5 arcseconds
flags & 0x0000000001000000	=0	The object is not defined as "too large" to be measured by SDSS pipeline
flags & 0x0000000000008000	=0	The object is not defined as "bad radial" by SDSS pipeline
flags & 0x0000000000000400	=0	The Petrosian radius is not beyond the last point in the radial profile
flags & 0x0000000000400000	=0	The sky is not defined as "bad sky" by SDSS pipeline
flags & 0x0000000000000004	=0	The object is not too close to the edge of the frame
flags & 0x0000002000000000	=0	Peaks were not too close
flags & 0x0000000000000200	=0	Only one Petrosian radius was found
flags & 0x0001000000000000	=0	No child of the object had too few good detection

by 0.01 arcseconds per pixel, and the image was re-downloaded and tested again until the object fitted the frame. The 25% threshold is used since in some cases parts of the galaxy or background objects such as stars can be located on the edge of the image, but since these objects are expected to be small no more than 25% of the edge pixels are expected to be bright. The simple method ensured that the entire object was inside the frame, surrounded by background sky.
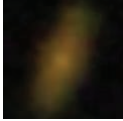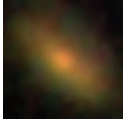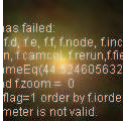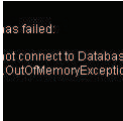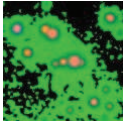
Although all images that were downloaded were images of objects flagged as galaxies by SDSS, some of the images contained artifacts, errors, or were actually images of astronomical objects that were not galaxies as shown in Table 2. Images of these non-galaxy objects can be identified by the image size, as the artifacts, saturated pixels, and error messages changed the compression factor of the im-

age to result in a much larger or much smaller file than a "clean" galaxy image. In some cases the image failed to download due to a server error, and since the error message is typed on the image the file becomes larger. By observing 200 images and their file sizes, we set a minimum threshold of 2000 bytes and an upper threshold of 6000 bytes. Based on our observations, objects with file size outside that range were not actual astronomical objects. Therefore, these images were removed from the catalog, resulting in the sacrifice of 27,992 images. Table 2 shows several examples of images and their file sizes.

2.1. Galaxy image analysis

The image analysis method used to classify the images is Wndchrm (Shamir et al. 2008, 2010, 2013b), that first computes 2885 numerical descriptors from each image such as textures (Haralick et al.

4

Table 2: Galaxy images and non-galaxy images in SDSS. All images were flagged by SDSS as galaxies, although some are clearly artifacts. The file size provides a simple mechanism to remove images of artifacts from the dataset

SDSS DR8 ID	File size (KB)	Image
1237645879037067801	3.13	
1237649743943369548	3.34	
1237671261735354668	3.47	
1237646012176400498	16.4	
1237645941290369211	8.96	
1237649769179578519	11.3	
1237672815434465362	8.5	
1237651752387411988	1.57	

statistical distribution of the pixel intensities (Hadjidemetriou et al. 2001), polynomial decomposition of the image (Teague 1980), and fractal features (Wu et al. 1992). These features are extracted from the raw pixels, as well as the image transforms and multi-order image transforms (Shamir et al. 2010). A complete and detailed description of the set of numerical image content descriptors and comprehensive performance analysis is available in (Shamir et al. 2008; Orlov et al. 2008; Shamir et al. 2010, 2009). The source code is also publicly available (Shamir et al. 2013b), and can be downloaded at `http://vfacstaff.ltu.edu/lshamir/downloads/ImageClassifier`.

In particular, Wndchrm has been proven to be informative for analysis of galaxy morphology, and was found useful for tasks such as galaxy classification (Shamir 2009; Kuminski et al. 2014), unsupervised analysis of galaxy images (Schutter and Shamir 2015; Shamir et al. 2013a), and automatic detection of peculiar galaxies (Shamir 2012; Shamir and Wallin 2014; Shamir et al. 2014).

Previous experiments have shown that Wndchrm achieves ∼95% of classification accuracy when separating spiral and elliptical galaxies (Shamir 2009), which outperforms methods such as Gini (Abraham et al. 2003) and CAS (Conselice 2003) as described in (Shamir 2009). Its performance is favorably comparable to the performance reported by using a Neural Network classifier (Banerji et al. 2010), and is also higher than methods based on the data provided by SDSS photometry pipeline (Yamauchi et al. 2005). The accuracy can also be improved when ignoring

1973; Tamura et al. 1978), edges (Prewitt 1970), shapes (Orlov et al. 2008),

classifications with lower certainty (Shamir 2009). A detailed description and experimental results of Wndchrm classification to spiral and elliptical galaxies is available in (Shamir 2009).

The training set with which the classifier was trained included 150 galaxies classified manually as spiral galaxies, and 150 galaxies classified manually as elliptical galaxies as was done in (Shamir 2009). The training images were first selected by manually choosing 80 galaxy images from each morphological type, and then using a classifier trained with these samples to classify a dataset of 1000 galaxy images. The classified images were inspected manually, and when a misclassified galaxy was noticed the process was stopped, the misclassified galaxy was added to the correct class in the training set, and the classifier was re-trained using the new training set. The process was repeated until no misclassified galaxies were observed.

The distribution of the Petrosian radius measured on the r band, r magnitude, and photometric redshift of the training galaxies are displayed in Figure 1.

As a classifier, Wndchrm does not just predict the class of a given sample, but also provides the user with the certainty of the classification to each class as a value within the interval [0,1] (Shamir et al. 2008, 2010). For instance, certainty of 1 means a high certainty of the classification of the object to that class, while if the certainty of the classification is 0.7 there is a higher chance that the classification is incorrect. The sum of certainties of a sample to belong in all classes is always 1.0 (Shamir et al. 2008; Shamir 2009). These computed certainties can be used in a similar fashion to the degree of agreement between the voters in a citizen science campaign. Therefore, thresholding these certainty values can be used to provide subset catalogs with different levels of accuracies, such that galaxies that do not meet the threshold can be considered "unknown". That practice allows controlling the accuracy/size trade-off, as will be discussed in Section 4.

3. Catalog of broad galaxy morphology

The catalog of galaxies is available for download at http://vfacstaff.ltu.edu/lshamir/data/morph_catalog in the form of a CSV (comma separated values) file. The catalog provides information about 2,911,899 galaxies. The information for each object includes the DR8 ID, right ascension, declination, certainty of the galaxy being elliptical, and the certainty of the galaxy being spiral as computed by the image analysis method described in Section 2.1.

The DR8 ID can be used as a handle to extract all other photometric or other information of interest about the galaxies through CAS.

The distribution of the size, r magnitude, and redshift (of objects with spectra) of the training galaxies are displayed in Figure 2.

3.1. Morphological catalog of SDSS objects with spectra

In a similar way we also compiled a catalog of all objects with spectra in DR8. That catalog contains for each object the specObjID, right ascension, declination, z, z error, certainty of classification to ellipti-
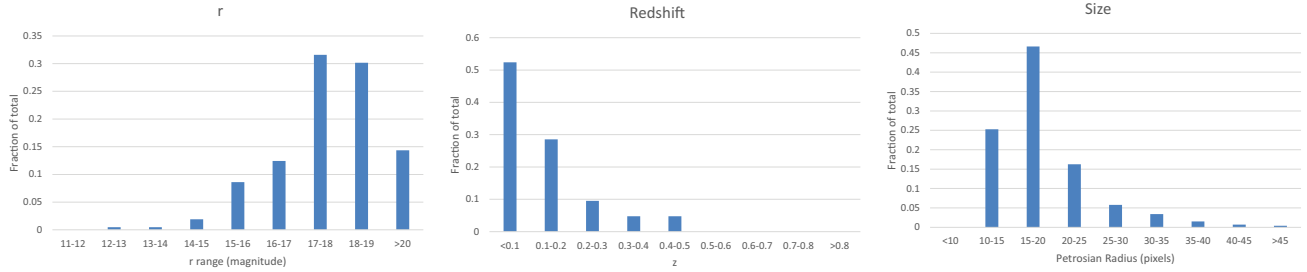
6

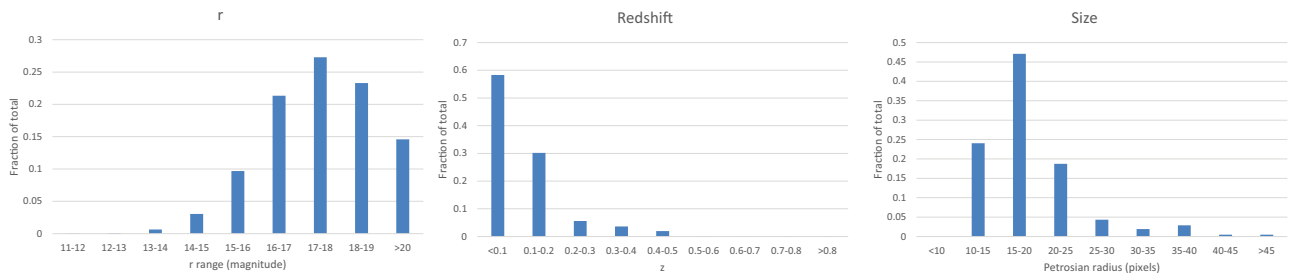Fig. 1.— Histograms of the r magnitude, Petrosian radius, and photometric redshift of the training galaxies.

Fig. 2.— Histograms of the r magnitude, Petrosian radius, and redshift of the galaxies in the catalog.

cal, certainty of classification to spiral, and certainty of classification to a star.

The catalog includes the morphological analysis of 2,638,823 astronomical objects with spectra. Unlike the catalog described in Section 3, the objects were not filtered by their size, so the catalog also contains the classification of small or faint objects that their SDSS image does not provide sufficient information to make a clear classification of the morphology.

The distribution of the size, r magnitude, and redshift of the training galaxies are displayed in Figure 3.

Because the catalog also includes many objects that are not galaxies, it contains a much larger number of smaller objects. SDSS photometry pipeline provides an automatic separation between stars and galaxies by the difference between the cmodel and PSF magnitude, but that separation is not perfect (Stoughton et al. 2002). By analyzing all objects with spectra, the catalog includes objects that could have been misclassified as stars, as well as providing information about objects that were not classified by SDSS photometry pipeline as galaxies. The classification in this catalog is based on the visual cues, and can be used in combination with SDSS PhotObjAll.

4. Evaluation

To evaluate the accuracy of the catalog we compared the classifications of the galaxies in the catalog to the manual clas-
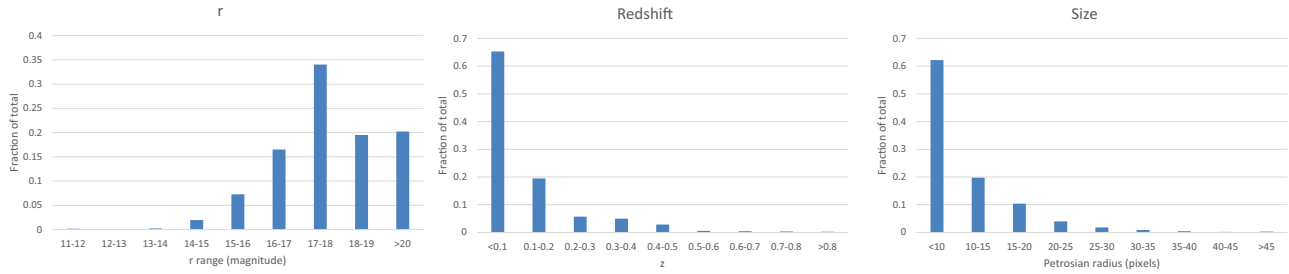
Fig. 3.— Histograms of the r magnitude, size, and redshift of the objects in the catalog.

sifications of the galaxies in the Galaxy Zoo catalog (Lintott et al. 2011). As a crowdsouring-based project it is possible that the classifications of the citizen scientists might in some cases disagree with the morphology of the object determined by a professional astronomer. However, the Galaxy Zoo catalog is based on manual classification of a large number of citizen scientists, applies bias correction mechanisms, and has been inspected for its accuracy (Lintott et al. 2011), making it suitable to provide a ground truth baseline to test the accuracy of computer-generated catalogs. Although Galaxy Zoo can theoretically contain errors, the number of misclassifications in the debiased "superclean" dataset is expected to be low. **The Galaxy Zoo debiased dataset is corrected for the redshift effect, and therefore the probability of a galaxy to be manually classified as elliptical is not expected to increase as the redshift gets higher**.

Since the number of galaxies analyzed for this catalog is far higher than the number of galaxies in Galaxy Zoo, it is clear that not all galaxies in the catalog are also analyzed by Galaxy Zoo. However, given the large number of galaxies in both datasets, even a mild overlap can provide

a sufficient number of galaxies to allow the analysis of the consistency of the classifications across the two datasets.

The total number of galaxies in the catalog that were also classified by Galaxy Zoo is 229,271. That includes all galaxies that were classified, and clearly some of these galaxies were not classified in high certainty that makes them suitable to be used as ground truth. Therefore, we used the Galaxy Zoo galaxies that were classified as "superclean", and were corrected for red-shift bias (Lintott et al. 2011). The number of galaxies in the catalog that were also classified by Galaxy Zoo with debiased "superclean" classification is 45,377.

Figure 4 shows the agreement between the computer classification and the Galaxy Zoo "superclean" manual classification for different levels of computer classification certainty.

As the figure shows, when the computer classification is a spiral galaxy with certainty higher than 0.6, the rate of agreement between the catalog and Galaxy Zoo superclean debiased data is ∼98.7%. That is, of all galaxies classified as spiral with certainty higher than 0.6 and were also classified by Galaxy Zoo superclean dataset, ∼98.7% were also classified as spi-
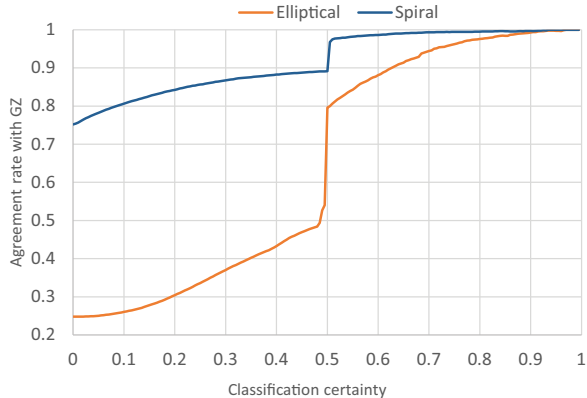
8

Fig. 4.— Rate of agreement with galaxies classified by Galaxy Zoo "superclean" debiased dataset for different ranges of computed classification certainties

ral in Galaxy Zoo superclean catalog.

When the certainty of a spiral galaxy classification is higher than 0.625 the agreement with Galaxy Zoo superclean data is ∼99%, and the rate of agreement is ∼98% when the certainty level is 0.54 or higher. These results show that when setting the certainty threshold to 0.54, the accuracy of the classification is very similar to the classifications of Galaxy Zoo superclean data.

Similarly, when the classification is elliptical galaxy in certainty of 0.8 or higher, ∼97.7% of the same galaxies that are also classified by Galaxy Zoo superclean dataset are also classified as elliptical. Expectedly, the accuracy leaps when the threshold reaches 0.5, which is the threshold beyond which a galaxy is considered to be more likely to be spiral (when spiral certainty is greater than 0.5) or elliptical (when elliptical classification certainty is greater than 0.5).

The figure also shows that the dis-

agreement rate is higher for the elliptical galaxies compared to the spiral galaxies. That difference can be explained by the higher sensitivity of the human eye to spiral galaxies compared to elliptical galaxies (Dojcsak and Shamir 2014). That is, when the spiral features of the galaxy are visible it is easier to classify the galaxy as spiral by manual observation. However, when the spirality of the galaxy is not easily noticeable it can mean that the galaxy is indeed not spiral, or that the image resolution is not high enough to identify the spirality. Therefore, it is expected that more galaxies classified as elliptical are actually spiral galaxies, compared to galaxies classified as spiral that are actually elliptical (Dojcsak and Shamir 2014).

Figure 5 shows the number of galaxies in the catalog classified as spiral or elliptical above different levels of certainty. Expectedly, when the certainty threshold is zero all galaxies meet the threshold, and the number of galaxies that meet the threshold gets smaller as the threshold increases. At certainty level 0.5 a sharp drop is observed, which corresponds to the drop in Figure 4.

Agreement rate of 98% with Galaxy Zoo "superclean" catalog is achieved when the certainty level for spiral galaxy classification is 0.54 or higher, and the certainty level for elliptical galaxies is 0.8 or higher. Therefore, the catalog contains ∼600,000 elliptical galaxies and ∼900,000 spiral galaxies that have 98% accuracy when considering Galaxy Zoo "superclean" data as ground truth. Since the computer analysis works differently from the human brain, the galaxies classified manually as "superclean" are not necessarily easier to
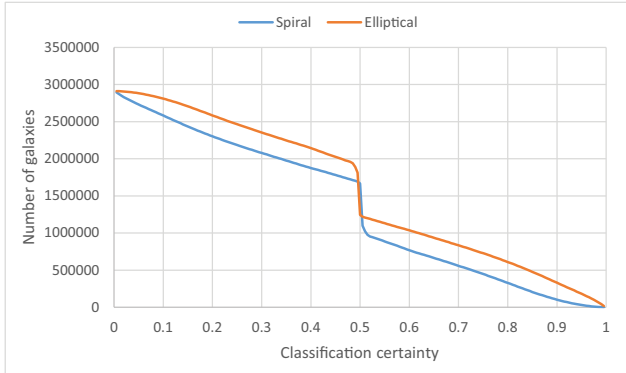
Fig. 5.— Number of spiral and elliptical galaxies classified above different certainty levels.

classify by the computer compared to other galaxies that are not "superclean". The comparison between human certainty and computer certainty is described in Section 5.

4.1. Evaluation of the catalog of objects with spectra

Similarly, we also evaluated the accuracy of the catalog of objects with spectra by using Galaxy Zoo debiased "superclean" data as ground truth. The number of objects in the catalog that were also classified by Galaxy Zoo is 620,529. Among these galaxies, 59,984 were classified with "superclean" accuracy. The high overlap between the objects with spectra and Galaxy Zoo is because Galaxy Zoo intentionally selected objects classified as galaxies by SDSS pipeline and also had spectroscopy information. Objects identified as stars, on the other hand, were not selected for Galaxy Zoo classification since Galaxy Zoo focused on the morphology on galaxies, and therefore asking the citizen scientists to classify the morphology of ob-

jects identified by SDSS as stars would be an inefficient use of resources.

Figure 6 shows the agreement between the Galaxy Zoo classification of these galaxies and the computer classification for different levels of computer classification certainty. Since the catalog of objects with spectra also contains small or faint objects with morphology that is more difficult to identify, the agreement between the two catalogs is lower compared to the catalog of objects without spectra.
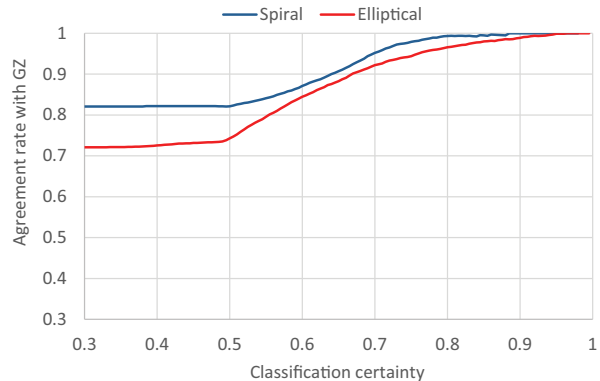

Fig. 6.— Rate of agreement between the catalog of objects with spectra and Galaxy Zoo "superclean" debiased dataset

The number of galaxies classified as spiral with different thresholds of classification certainty are displayed in Figure 7. As the figure shows, the number of objects identified as spiral galaxies is much lower in the catalog of objects with spectra. That can be explained by the fact that the objects in that catalog were not filtered by their size, so small objects with no apparent spirality were more common in that catalog.

Figure 8 shows the number of Galaxy Zoo galaxies above different agreement rates. Expectedly, a higher agreement rate
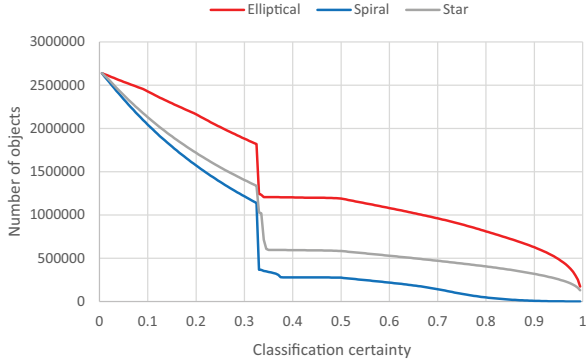
Fig. 7.— Number of spiral galaxies, elliptical galaxies, and stars with spectra classified above different certainty levels.

leads to fewer galaxies, but the decline is more consistent compared to the computer classification. The decline in the number of galaxies that meets the agreement thresholds starts from agreement rate of 0 because the classifications are not compared to external ground truth, and therefore majority of the votes does not lead to higher consistency with a certain standard of correct classifications.
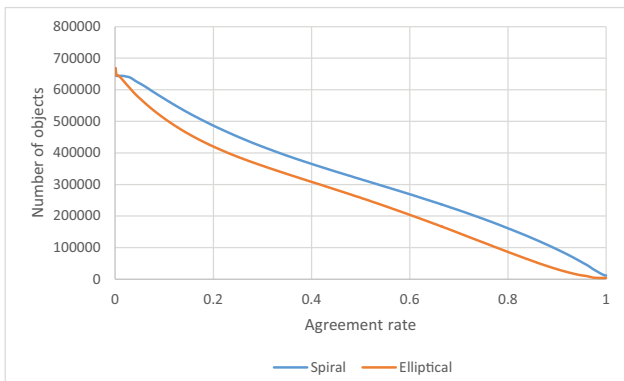
Fig. 8.— Number of spiral and elliptical galaxies with spectra above different agreement rates in Galaxy Zoo debiased dataset

5. Correlation between the computer and manual classification

When comparing machine and human classification of galaxies, more accurate human classification leads to higher consistency with machine learning methods performing the same task (Shamir et al. 2015). However, the computer analysis is clearly not an attempt to mimic the cognitive processes activated inside the human brain, but merely an attempt to produce similar output as the output that would have been provided by careful manual classification of the images. Therefore, it is not clear whether galaxies that are easy to classify manually are also easier to classify by the computer. Figure 9 shows the Pearson correlation between the galaxy zoo agreement rate and the certainty of the computer classification of the same galaxies.
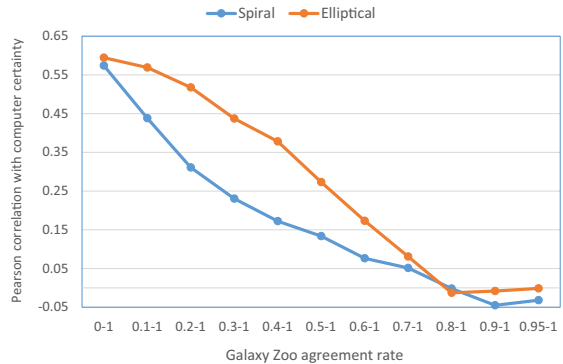

Fig. 9.— Pearson correlation between the manual and computer certainty for different ranges of GZ agreement rate.

The graph shows that for galaxies with high agreement rate the correlation between the citizen scientist agreement rate, which can be used as an indication for the easiness of the manual classification, has merely weak correlation with the certainty

of the classification computed by the machine learning method. The correlation becomes higher when the agreement threshold gets lower, and that can be explained by the fact that the classification of most galaxies in the two catalogs agree, leading to a correlation between the classifications.

It also shows that the correlation gets weaker when the galaxies are classified with higher agreement by the galaxy zoo participants, showing that galaxies that are easier to classify manually are not necessarily classified by the computer with higher certainty. In fact, for galaxies classified by the participants with agreement rate of 80% or higher there is virtually no correlation with the certainty of the classification computed by the machine learning method.

Figure 10 shows the same analysis such that all ranges are equal, so that no bias caused by the broader ranges of the galaxies with lower agreement rate can affect the computed correlation. As in Figure 9, the correlation with the computer certainty gets lower when the galaxies are easier to classify manually.

The observation that galaxies that are easier to classify manually are not necessarily easier to classify automatically is also reflected by the average certainties of the automatic classifications. Figure 11 shows the average certainty of the computer classifications for different ranges of citizen scientist agreement rates.

As the figure shows, when the agreement rate of the citizen scientists is higher than 80% the average certainty of the automatic classification of the galaxies does not increase with the agreement rate of the citizen scientists.
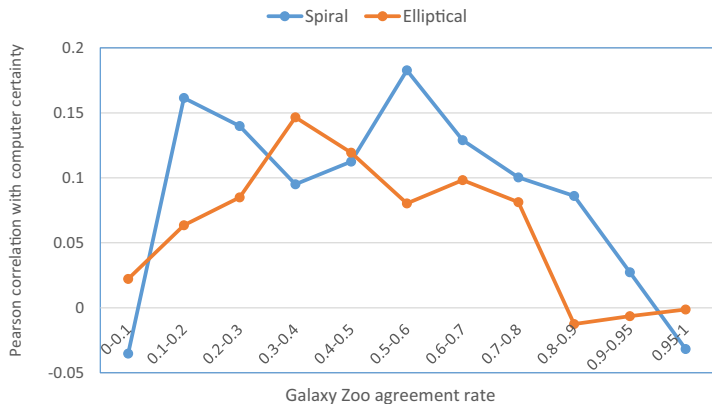

Fig. 10.— Pearson correlation between the manual and computer certainty for different ranges of GZ agreement rates.

6. Completeness of the catalog

As discussed in Sections 4, achieving a higher classification accuracy requires the sacrifice of many of the galaxies that their classification does not meet the certainty threshold. Therefore, given a certain threshold, the classifications of some galaxies do not meet that threshold, leading to an incomplete catalog. For instance, setting a threshold of 98% might lead to a relatively accurate catalog as described in Section 4, but at the same time will not include many galaxies that belong in that morphological type but do not meet the certainty threshold. As shown by Figure 5, when the classification threshold is higher, the number of galaxies that their classification certainty meets the threshold gets substantially lower, and therefore many galaxies that might belong in that morphological type are left outside of the threshold cutoff.

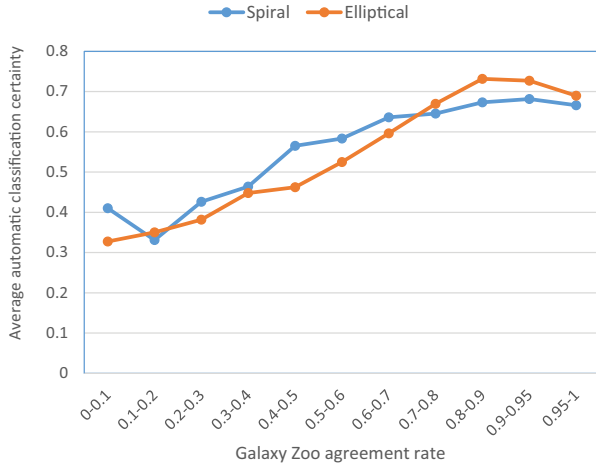Figure 12 shows the amount of galaxies that were classified by Galaxy Zoo as

Fig. 11.— Average computer classification certainty in different ranges of GZ agreement rates.

Fig. 12.— The completeness of the catalog compared to debiased "superclean" Galaxy Zoo classifications.

debiased ''superclean'', that were classified by the algorithm with certainty lower than different levels of certainty thresholds. For instance, ∼72% of the galaxies classified by Galaxy Zoo as ''superclean'' spiral galaxies were not classified by the algorithm with certainty higher than 0.8, showing that improving the classification accuracy by increasing the certainty threshold leads to the sacrifice of galaxies that are of the same morphological type, but the certainty of their automatic classification does not meet the threshold.

The figure shows that when assuming all debiased "superclean" Galaxy Zoo galaxies are annotated correctly, the classification accuracy of ∼98% of the spiral galaxies achieved when the threshold is set to 0.54 as shown in Figure 4 is achieved when sacrificing ∼30% of the galaxies that are actually spiral.

Figure 13 shows the number of debiased ''superclean'' Galaxy Zoo galaxies that were also classified by the automatic

classifier in different certainty thresholds. Clearly, the number of ''superclean'' galaxies that were also identified by the algorithm decreases as the threshold gets higher.

Expectedly, the number of galaxies that do not meet the threshold drops sharply when the certainty threshold is below 0.5. However, a certain number of galaxies that were classified by Galaxy Zoo as "superclean" were misclassified by the algorithm. As Figure 12 shows, when the certainty is below 0.3 the number of "superclean" galaxies that their classification disagrees with the classification of the algorithm is very low, but it is not zero. Four debiased "superclean" elliptical galaxies were classified by the algorithm as elliptical with certainty of 0.3 or lower, and 420 debiased "superclean" spiral galaxies were classified by the algorithm as spiral with certainty lower than 0.3. Table 3 shows the galaxies that were classified by Galaxy Zoo as "su-
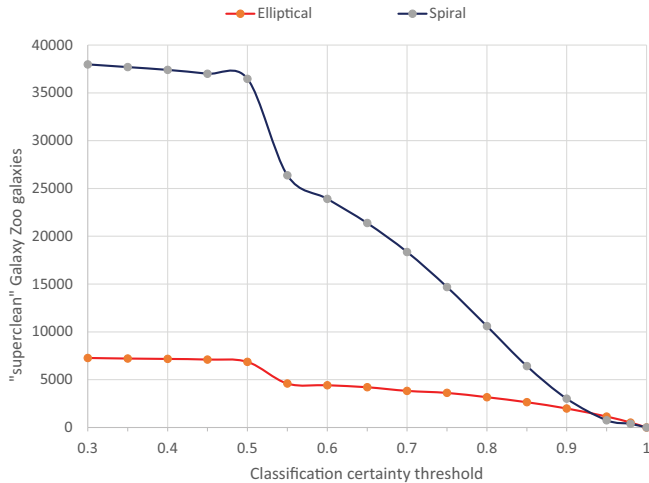
13

Fig. 13.— The number of debiased "super-clean" Galaxy Zoo classifications that were also classified by the algorithm in different certainty thresholds.

perclean" elliptical but the algorithm classified them as spiral galaxies with certainty higher than 0.7 (and therefore as elliptical with certainty lower than 0.3).

None of these galaxies has visible spiral arms, and therefore the misclassification can be attributed to noise in the pattern recognition systems. Since machine vision image recognition systems are often not fully accurate, it can be expected that some galaxies can be misclassified, as also discussed in Section 4. All four galaxies have a clear nucleus that might also lead to higher similarity to spiral galaxies, but the galaxies in Table 3 show that in some (rare) cases elliptical galaxies can be misclassified as spirals with relatively high certainty.

Table 4 shows example galaxies that were classified automatically as elliptical with certainty higher than 0.7, but classified as "superclean" spiral galaxies by

Table 3: Galaxies classified as debiased "superclean" elliptical by Galaxy Zoo but as spiral by the automatic classifier

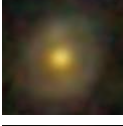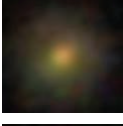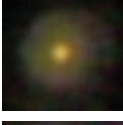| SDSS DR8 ID	Computed elliptical certainty	Image
1237650372096688281	0.069772	
1237663789575438589	0.092223	
1237664093439328427	0.093762	
1237671260125200613	0.084093	

Galaxy Zoo.

Galaxy ID 1237654610682773622 does not show clear spiral arms, demonstrating that galaxies classified by Galaxy Zoo as "superclean" might also contain misclassified galaxies. The other galaxies are clearly not elliptical, and provide examples of galaxies that were misclassified by the algorithm. These galaxies normally have dimmer arms compared to their relatively bright nucleus. As discussed in Section 4, the algorithm is not in perfect agreement with manual analysis, but galaxies that are misclassified despite being classified with a high certainty are relatively rare.

Figure 14 shows the distribution of the galaxies in different ranges of r magnitude, Petrosian radius (measured on the r band) and redshift. The figure shows the distri-

Table 4: Galaxies classified as debiased "superclean" spiral by Galaxy Zoo but as elliptical galaxies by the algorithm

SDSS DR8 ID	Computed spiral certainty	Image
1237651192424366347	0.142726	
1237651192429412678	0.290797	
1237648721759371433	0.190888	
1237658203978203292	0.120782	
1237657219874619557	0.241804	
1237654610682773622	0.189278	
1237654604795674718	0.272735	
1237651737381306386	0.115425	

bution of Galaxy Zoo galaxies, Galaxy Zoo "superclean" galaxies, galaxies that were classified with certainty of 0.8 or higher for elliptical galaxies, and 0.54 or higher for spiral galaxies, which is in 98% agreement with Galaxy Zoo debiased "superclean" catalog as described in Section 4, and galaxies that their classification certainty did not meet that criteria.

Expectedly, "superclean" galaxies are more prevalent in lower redshift, lower r, and larger galaxy size. The computer classification is also more accurate for brighter objects, and when the redshift is lower. Brighter objects provide more visual details that can be analyzed by the computer, and the correlation between r and z leads to difference in z distribution between classified and unclassified galaxies. However, the computer classification is less sensitive to size, probably because the galaxies were initially selected by their Petrosian radius and then downscaled to fit the frame as described in Section 2, providing a dataset that does not contain small object.

Galaxies with higher redshift are expected to be smaller and fainter, and therefore their morphology might be more difficult to classify. Since the galaxies are selected by their size and analyzed by their morphology, the catalog can also be used to identify galaxies that SDSS spectroscopy pipeline assigned with high redshift, and have identifiable visual spiral morphology. For instance, 161 objects were classified as spiral galaxies with certainty higher than 0.54, and were associated in SDSS DR8 with spectroscopic objects with z higher than 0.4. Table 5 shows examples of these galaxies.

7. Computational complexity

As mentioned above, downloading the images of the catalog of objects without spectra from CAS was completed within
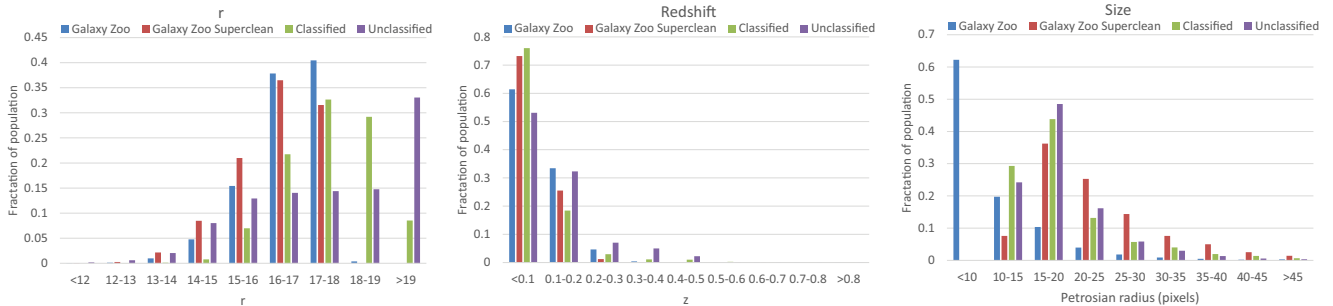
Fig. 14.— Distribution of the r magnitude, Petrosian radius (r band), and redshift among Galaxy Zoo galaxies classified as debiased "superclean", galaxies of the catalog that were classified at the accuracy level of 98% agreement with the debiased "superclean" Galaxy Zoo galaxies, and galaxies that their classification did not meet that level.

~25 days. The analysis of the galaxies is done in two stages. The first is computing the numerical image content descriptors of the galaxy, and the second is the classification of the galaxy based on its numerical image content descriptors.

Computing the numerical image content descriptors of one galaxy using a single Intel Core-i7 processor takes ~45 seconds. Wndchrm can be easily parallelized so that each different instances running on different cores process different galaxies (Shamir et al. 2008). Using 128 cores, the numerical image content descriptors of the galaxies were computed in ~12 days. The classification of all galaxies was done using a single core, and lasted ~75 hours.

8. Conclusion

Here we describe a catalog of ~3M galaxies classified by their broad morphological type. The objects in the catalog can be selected based on the certainty of the classification to control the quantity-quality trade-off. The results show that ~1.5M galaxies are classified with a de-

gree of certainty that matches Galaxy Zoo "superclean" classifications in 98% of the cases, making the catalog by far the largest of its kind.

Due to the imperfectness of the classification algorithm, higher accuracy is achieved by sacrificing some of the galaxies that do not meet a certain threshold of classification certainty, and therefore the catalog is clearly incomplete. However, given the increasing gap between the power of astronomical data collection devices and the automatic analysis tools that can turn these data into knowledge, sacrificing some of the data can still result in very large catalogs.

That also demonstrates the ability of computer programs to create clean morphological catalogs containing millions of astronomical objects. As current and future digital sky surveys have been becoming increasingly more powerful, automatic methods for analyzing galaxies are required to analyze these data. Digital sky surveys such as the Large Synoptic Sky Survey (LSST) are expected to acquire images of billions of galaxies, rein-

forcing the need to automatically analyze the morphology of these objects and allow scientific discoveries using these data. Therefore, computer vision methods that can produce clean catalogs of morphological descriptors of astronomical objects will become essential to fully utilized the large databases created by digital sky surveys.

9. Acknowledgments

REFERENCES

Abell, P. A., Allison, J., Anderson, S. F., Andrew, J. R., Angel, J. R. P., Armus, L., Arnett, D., Asztalos, S., Axelrod, T. S., Bailey, S., et al. (2009). Lsst science book, version 2.0. *arXiv preprint arXiv:0912.0201*.

Abraham, R. G., Van Den Bergh, S., and Nair, P. (2003). A new approach to galaxy morphology. i. analysis of the sloan digital sky survey early data release. *The Astrophysical Journal*, 588(1):218.

Baillard, A., Bertin, E., De Lapparent, V., Fouqué, P., Arnouts, S., Mellier, Y., Pelló, R., Leborgne, J.-F., Prugniel, P., Makarov, D., et al. (2011). The efigi catalogue of 4458 nearby galaxies with detailed morphology. *Astronomy & Astrophysics*, 532:A74.

Banerji, M., Lahav, O., Lintott, C. J., Abdalla, F. B., Schawinski, K., Bamford, S. P., Andreescu, D., Murray, P., Raddick, M. J., Slosar, A., et al. (2010). Galaxy zoo: reproducing galaxy morphologies via machine learning. *Monthly Notices of the Royal Astronomical Society*, 406(1):342–353.

Borne, K. (2013). Virtual observatories, data mining, and astroinformatics. In *Planets, Stars and Stellar Systems*, pages 403–443. Springer.
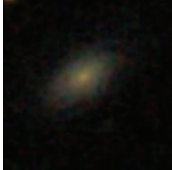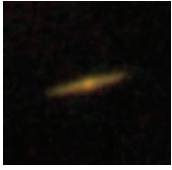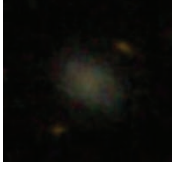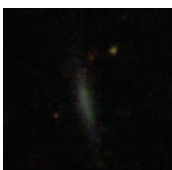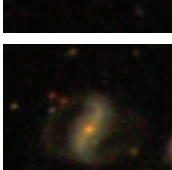
Conselice, C. J. (2003). The relationship between stellar light distributions of galaxies and their formation histories. *The Astrophysical Journal Supplement Series*, 147(1):1.

Davis, D. R. and Hayes, W. B. (2014). Sparcfire: Scalable automated detection of spiral galaxy arm segments. *The Astrophysical Journal*, 790(2):87.

Dieleman, S., Willett, K. W., and Dambre, J. (2015). Rotation-invariant convolutional neural networks for galaxy morphology prediction. *Monthly Notices of the Royal Astronomical Society*, 450(2):1441–1459.

Djorgovski, S. G., Mahabal, A., Drake, A., Graham, M., and Donalek, C. (2013). Sky surveys. In *Planets, Stars and Stellar Systems*, pages 223–281. Springer.

Dojcsak, L. and Shamir, L. (2014). Quantitative analysis of spirality in elliptical galaxies. *New Astronomy*, 28:1–8.

Hadjidemetriou, E., Grossberg, M. D., and Nayar, S. K. (2001). Spatial information in multiresolution histograms. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 1, pages 702–709. IEEE.

Haralick, R. M., Shanmugam, K., and Dinstein, I. H. (1973). Textural features for image classification. *IEEE Transactions on Systems, Man and Cybernetics*, (6):610–621.

Huertas-Company, M., Aguerri, J., Bernardi, M., Mei, S., and Almeida, J. S. (2010). Revisiting the hubble sequence in the sdss dr7 spectroscopic sample: a publicly available bayesian automated classification. *arXiv preprint arXiv:1010.3018*.

Huertas-Company, M., Gravet, R., Cabrera-Vives, G., Pérez-González, P., Kartaltepe, J., Barro, G., Bernardi, M., Mei, S., Shankar, F., Dimauro, P., et al. (2015a). A catalog of visual-like morphologies in the 5 candels fields using deep-learning. *arXiv preprint arXiv:1509.05429*.

Huertas-Company, M., Pérez-González, P. G., Mei, S., Shankar, F., Bernardi, M., Daddi, E., Barro, G., Cabrera-Vives, G., Cattaneo, A., Dimauro, P., et al. (2015b). The morphologies of massive galaxies from z~ 3-witnessing the 2 channels of bulge growth. *arXiv preprint arXiv:1506.03084*.

Huertas-Company, M., Tasca, L., Rouan, D., Pelat, D., Kneib, J., Le Fevre, O., Capak, P., Kartaltepe, J., Koekemoer, A., Mccracken, H., et al. (2009). A robust morphological classification of high-redshift galaxies using support vector machines on seeing limited images. *Astronomy and Astrophysics*, 497(3):743.

Kuminski, E., George, J., Wallin, J., and Shamir, L. (2014). Combining human and machine learning for morphological analysis of galaxy images. *Publications of the Astronomical Society of the Pacific*, 126(944):959–967.

Lintott, C., Schawinski, K., Bamford, S., Slosar, A., Land, K., Thomas, D., Edmondson, E., Masters, K., Nichol, R. C.,

Raddick, M. J., et al. (2011). Galaxy zoo 1: data release of morphological classifications for nearly 900 000 galaxies. *Monthly Notices of the Royal Astronomical Society*, 410(1):166–178.

Lintott, C. J., Schawinski, K., Slosar, A., Land, K., Bamford, S., Thomas, D., Raddick, M. J., Nichol, R. C., Szalay, A., Andreescu, D., et al. (2008). Galaxy zoo: morphologies derived from visual inspection of galaxies from the sloan digital sky survey. *Monthly Notices of the Royal Astronomical Society*, 389(3):1179–1189.

Nair, P. B. and Abraham, R. G. (2010). A catalog of detailed visual morphological classifications for 14,034 galaxies in the sloan digital sky survey. *The Astrophysical Journal Supplement Series*, 186(2):427.

Orlov, N., Shamir, L., Macura, T., Johnston, J., Eckley, D. M., and Goldberg, I. G. (2008). Wnd-charm: Multipurpose image classification using compound image transforms. *Pattern Recognition Letters*, 29(11):1684–1693.

Peng, C. Y., Ho, L. C., Impey, C. D., and Rix, H.-W. (2002). Detailed structural decomposition of galaxy images. *The Astronomical Journal*, 124(1):266.

Prewitt, J. M. (1970). Object enhancement and extraction. *Picture Processing and Psychopictorics*, 10(1):15–19.

Schutter, A. and Shamir, L. (2015). Galaxy morphology-an unsupervised machine learning approach. *Astronomy and Computing*, 12:60–66.

Shamir, L. (2009). Automatic morphological classification of galaxy images. *Monthly Notices of the Royal Astronomical Society*, 399(3):1367–1372.

Shamir, L. (2011). Ganalyzer: A tool for automatic galaxy image analysis. *The Astrophysical Journal*, 736(2):141.

Shamir, L. (2012). Automatic detection of peculiar galaxies in large datasets of galaxy images. *Journal of Computational Science*, 3(3):181–189.

Shamir, L., Diamond, D., and Wallin, J. (2015). Leveraging pattern recognition consistency estimation for crowdsourcing data analysis. *IEEE Transactions on Human-Machine Systems*, In Press.

Shamir, L., Holincheck, A., and Wallin, J. (2013a). Automatic quantitative morphological analysis of interacting galaxies. *Astronomy and Computing*, 2:67–73.

Shamir, L., Ling, S. M., Scott, W. W., Bos, A., Orlov, N., Macura, T. J., Eckley, D. M., Ferrucci, L., and Goldberg, I. G. (2009). Knee x-ray image analysis method for automated detection of osteoarthritis. *IEEE Transactions on Biomedical Engineering*, 56(2):407–415.

Shamir, L., Macura, T., Orlov, N., Eckley, D. M., and Goldberg, I. G. (2010). Impressionism, expressionism, surrealism: Automated recognition of painters and schools of art. *ACM Transactions on Applied Perception*, 7(2):8.

Shamir, L., Manning, S., and Wallin, J. (2014). Chloe: A tool for automatic detection of peculiar galaxies. *As-*

trophysics Source Code Library, page ascl:1409.008.

Shamir, L., Orlov, N., Eckley, D. M., Macura, T., Johnston, J., and Goldberg, I. (2013b). Wnd-charm: Multipurpose image classifier. *Astrophysics Source Code Library*, page ascl:1312.002.

Shamir, L., Orlov, N., Eckley, D. M., Macura, T., Johnston, J., and Goldberg, I. G. (2008). Wndchrm – an open source utility for biological image analysis. *Source Code for Biology and Medicine*, 3:13.

Shamir, L. and Wallin, J. (2014). Automatic detection and quantitative assessment of peculiar galaxy pairs in sloan digital sky survey. *Monthly Notices of the Royal Astronomical Society*, 443(4):3528–3537.

Simard, L. (1999). Photometric redshifts and the luminosity-size relation of galaxies to z= 1. 1. In *Photometric Redshifts and the Detection of High Redshift Galaxies*, volume 191, page 325.

Simard, L., Mendel, J. T., Patton, D. R., Ellison, S. L., and McConnachie, A. W. (2011). A catalog of bulge+ disk decompositions and updated photometry for 1.12 million galaxies in the sloan digital sky survey. *The Astrophysical Journal Supplement Series*, 196(1):11.

Stoughton, C., Lupton, R. H., Bernardi, M., Blanton, M. R., Burles, S., Castander, F. J., Connolly, A., Eisenstein, D. J., Frieman, J. A., Hennessy, G., et al. (2002). Sloan digital sky survey:

early data release. *The Astronomical Journal*, 123(1):485.

Tamura, H., Mori, S., and Yamawaki, T. (1978). Textural features corresponding to visual perception. *IEEE Transactions on Systems, Man and Cybernetics*, 8(6):460–473.

Teague, M. R. (1980). Image analysis via the general theory of moments. *Journal of the Optical Society of America*, 70(8):920–930.

Willett, K. W., Lintott, C. J., Bamford, S. P., Masters, K. L., Simmons, B. D., Casteels, K. R., Edmondson, E. M., Fortson, L. F., Kaviraj, S., Keel, W. C., et al. (2013). Galaxy zoo 2: detailed morphological classifications for 304 122 galaxies from the sloan digital sky survey. *Monthly Notices of the Royal Astronomical Society*, page stt1458.

Wu, C.-M., Chen, Y.-C., and Hsieh, K.-S. (1992). Texture features for classification of ultrasonic liver images. *IEEE Transactions on Medical Imaging*, 11(2):141–152.

Yamauchi, C., Ichikawa, S.-i., Doi, M., Yasuda, N., Yagi, M., Fukugita, M., Okamura, S., Nakamura, O., Sekiguchi, M., and Goto, T. (2005). Morphological classification of galaxies using photometric parameters: the concentration index versus the coarseness parameter. *The Astronomical Journal*, 130(4):1545.

York, D. G., Adelman, J., Anderson Jr, J. E., Anderson, S. F., Annis, J., Bahcall, N. A., Bakken, J., Barkhouser, R., Bastian, S., Berman, E., et al. (2000).

The sloan digital sky survey: Technical summary. *The Astronomical Journal*, 120(3):1579.

Table 5: Example galaxies with computed spiral certainty higher than 0.54, and were associated with spectroscopic objects in SDSS DR8 with redshift higher than 0.4.

Number	SDSS DR8 ID	SDSS DR8 z	Image
1	1237663278465810600	0.889	
2	1237668602608615569	0.428	
3	1237671261743022362	1.044	
4	1237651272422850916	0.534	
5	1237648705126400194	0.908	
6	1237668297140404560	1.262	
7	1237665442603926125	0.998	
8	1237662530608562195	0.401	