Automated Detection of Galaxy Groups Through Probabilistic Hough Transform

Rafee T. Ibrahem[1](✉), Peter Tino[1], Richard J. Pearson[1],
Trevor J. Ponman[1], and Arif Babul[2]

[1] University of Birmingham, Birmingham B15 2TT, UK
{rti273,p.tino}@cs.bham.ac.uk, {richard,tjp}@star.sr.bham.ac.uk
[2] University of Victoria, Victoria, BC V8P 5C2, Canada
babul@uvic.ca

Abstract. Galaxy groups play a significant role in explaining the evolution of the universe. Given the amounts of available survey data, automated discovery of galaxy groups is of utmost interest. We introduce a novel methodology, based on probabilistic Hough transform, for finding galaxy groups embedded in a rich background. The model takes advantage of a typical signature pattern of galaxy groups known as "fingers-of-God". It also allows us to include prior astrophysical knowledge as an inherent part of the method. The proposed method is first tested in large scale controlled experiments with 2-D patterns and then verified on 3-D realistic mock data (comparing with the well-known friends-of-friends method used in astrophysics). The experiments suggest that our methodology is a promising new candidate for galaxy group finders developed within a machine learning framework.

Keywords: Pattern Recognition · Probabilistic Hough transform · Galaxy group finder

1 Introduction

In general, galaxies tend to expand away from one another. However, in certain regions of space there can be an overdensity of galaxies. This results in sufficiently strong gravitational field so that nearby galaxies cannot escape from one another and remain bound together. Galaxy groups play a significant role in explaining the evolution of the universe and measuring its baryonic content. They can also signify gravitational lenses and contribute to the estimation of cosmological parameters [11]. Last but not least, galaxy groups act as laboratories to study different types of galaxy group evolution [18]. Many big galaxy redshift surveys have been conducted to identify galaxy positions in the sky and the recession (line-of-sight) velocities. Given the amounts of available survey data, automated discovery of galaxy groups is of utmost interest to astrophysicists.

One of the common galaxy group finders (with redshift information) is the Friends-of-Friends algorithm (FOF) [7]. The groups are located based on spatial information, linking particles (galaxies) within a pre-specified linking size.

S. Arik et al. (Eds.): ICONIP 2015, Part III, LNCS 9491, pp. 323–331, 2015.
DOI: 10.1007/978-3-319-26555-1_37

Close-by linked pairs (friends) are further aggregated into groups (friends of friends). The linking size is specified according to typical overdensity of galaxies within groups. Several approaches have been proposed to determine the linking size and to measure local galaxy densities [3,14]. Other approaches to galaxy group finding have been based on probabilistic formulations, extending FOF, or including model-based analysis [4,8]. The probabilistic framework enables one to deal consistently with issues such as redshift distortion.

Galaxy groups exhibit a characteristic "fingers of God" (FOG) shape in the angular-Z (redshift) plots - a prolonged dense structure centered at the group position and oriented along the line-of-sight (LOS). We propose to take advantage of such group signatures and develop a dedicated form of model-based probabilistic Hough transform (PHTM). In general, the existing galaxy group finders have many free parameters that need to be carefully set before applying the analysis. This raises issues regarding generality of the results and stability of the calibration process. Hough transform based models have been shown effective in the detection of patterns of interest in cluttered scenes [10]. Probabilistic Hough transform formulation enables us to include explicitly prior expectations on the shape of interest (FOG) through the likelihood model and to treat the background noise consistently.

Hough transform ideas have already been used in astronomy in other contexts, e.g. detection of circular or arc-like forms typically indicative of gravitational lensing [6], identification of continuous gravitational wave signals [2], detection of radial structures on the solar corona [9], or cleaning of the SuperCOSMOS Sky Survey (SSS) from the foreground/background noise [16].

The paper has the following organization: After briefly describing the nature of the problem and the data in Sect. 2, our Probabilistic Hough Transform Method (PHTM) is introduced in Sect. 3. We present experimental results in Sect. 4 and conclude in Sect. 5.

2 The Problem of Galaxy Group Identification

The observer on the Earth surveys the universe on a certain patch of the sky identified through two angles - Right Ascension (RA) and Declination (Dec). Besides the spatial position on the sky (RA, Dec), the velocity of the object along the LOS can be deduced from the redshift Z. A typical example of the form of a galaxy survey is shown (as a 2-D slice) in Fig. 1 [1]. Some of the FOG prolonged features (patterns) along the LOS signifying the presence of galaxy groups are clearly visible (marked by red ellipses), some are masked by the background. The challenge is to detect patterns corresponding to the real galaxy groups (true positives), while reducing the detection of similar patterns formed by the fore/background and chance superposition (false positives).

Due to lack of space, we will only briefly outline generation of realistic data involving galaxy groups used in our experiments. The generation process is rather

[1] constructed based on a figure from [1].

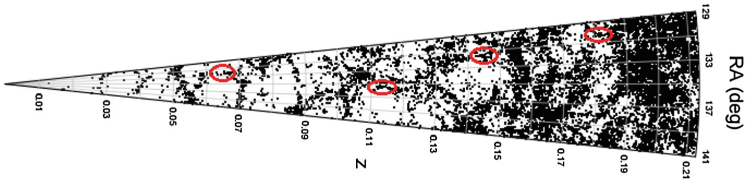


Fig. 1. 2-D slice from a volume of GAMA mock data: RA vs Z. Red elipses signify FOG features [1] (Color figure online).

involved and will be described in detail elsewhere[2]. To test our approach for close redshifts $Z \leq 0.1$, 3-D realistic data consists of two parts - galaxy groups themselves and fore/background galaxies. The key element of the data generation is generation of the individual galaxy groups (while carefully controlling for the extent of galaxy groups of given magnitudes at given Z). The groups are generated from a joint distribution consisting of a Gaussian distribution of dispersed projected velocities along the LOS and the radial distribution in the orthogonal complement of LOS formulated using the Navarro, Frenk and White (NFW) density profiles [12].

Before the methodology is demonstrated on realistic 3-D data, we will first test our method in a large set of controlled experiments in 2-D, where we control for the amount of background noise. In the 2-D setting the LOS direction is the y-axis and the galaxy groups are represented by points (galaxies) generated from Gaussian distributions elongated along the y-axis. In each group we generate 10–25 points from such Gaussian distributions. The background is generated from uniform distribution. The number of background points is determined as $T \cdot N_g$, where N_g is the number of galaxies in galaxy groups and T is a multiplicative factor in the range 5–30. In each setting there are 6 galaxy groups at fixed positions shown in Fig. 2a. A sample of 2D test data obtained with $T = 25$ is presented in 2b.

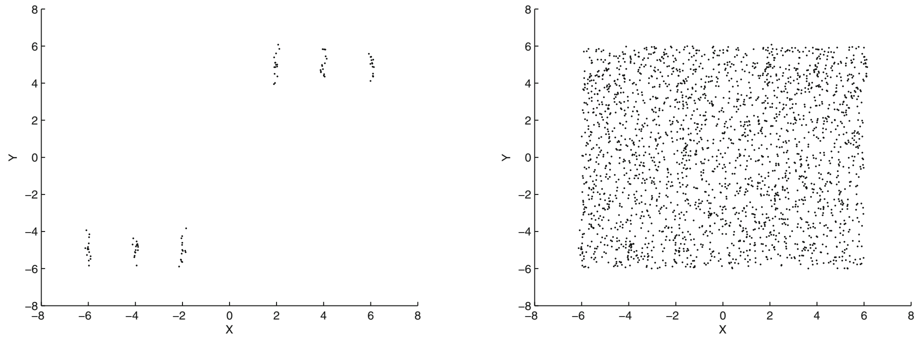
3 Probabilistic Hough Transform Galaxy Finder

Inspired by a probabilistic formulation of Hough Transform for co-expressed gene detection in 3-color cDNA arrays [17], we have developed a dedicated probabilistic Hough Transform method (PHTM) for galaxy group detection. We first introduce a simplified 2-D model to demonstrate the robustness of the PHTM approach and then introduce the full methodology operating in the 3-D realistic data, taking into account flux limit effects (more distant galaxies of the same intensity are less likely to be observed than closer ones).

3.1 Basic 2-D PHTM Group Finder

The search space is covered by a regular structure of G grid points. On each grid point, we position a noise model representing a possible galaxy group and

[2] codes will be available from www.cs.bham.ac.uk/~pxt/my.publ.html.



(a)

(b)

Fig. 2. 2D mock data: (a) Six galaxy groups at fixed positions; (b) galaxy groups with a background density equivalent to 25 times the number of galaxies in galaxy groups.

ask all observed galaxies to ascertain whether they are likely to have come from that group. Formally, for the i-th grid point (x_i, y_i) we have a Gaussian noise model centered at $\mu_i = (x_i, y_i)$ with axis-aligned (diagonal) covariance matrix $C = diag(k \cdot \sigma^2, \sigma^2)$. with variance along the y-axis σ^2 and variance along the x-axis $1/k$ times smaller, i.e. $k \cdot \sigma^2$ (we used $\sigma = 0.5$ and $k = 0.025$). The likelihood model for the i-th grid point is thus a multivariate Gaussian with mean μ_i and covariance C, $p(g|(x_i, y_i), C) = \mathcal{N}(\mu_i, C)$.

Given a galaxy g_q, $q = 1, 2, ..., N$, the degree to which it belongs to the possible group centered at the i-th grid point μ_i is quantified through posterior

$$P(i|g_q) = \frac{p(g_q|\mu_i, C) \cdot) P(i)}{\sum_{j=1}^{G} P(g_q|\mu_j, C)) \cdot P(j)}. \tag{1}$$

We assume no preferred positions for galaxy groups, i.e. flat prior $P(i) = 1/G$. The posterior can be interpreted as a ‘soft’ vote of the q-th galaxy for the possible galaxy group at position μ_i. The overall vote for the presence of galaxy group at μ_i is then obtained as a flat mixture of posteriors given by the observed galaxies:

$$H(x_i, y_j) = \frac{1}{N} \sum_{q=1}^{N} P(i|g_q). \tag{2}$$

Given a detection threshold $\Theta > 0$, the possible galaxy groups are detected as peaks above Θ in the $H(x_i, y_j)$ landscape. Note that high values of Θ will produce over-cautious conservative detections with a significant number of undetected true galaxy groups (false negatives). On the other hand, low Θ will lead to insignificant low peaks declared as group candidates (false positives).

3.2 Full 3-D PHTM Group Finder in Observational Cone

There are two principal modifications to be made to transform the fundamental model of Sect. 3.1 to the realistic case 3-D mock data (θ, β, Z), where θ and β

denote the RA and Dec, respectively. First, the noise model representing the idea of a galaxy group will be a 3-D Gaussian formulated in the corresponding Cartesian coordinate system (x, y, z) and elongated along the LOS (original axis Z in the cone). Second, in reality, due to the limited sensitivity of observational devices, more distant galaxies are less likely detected than the comparable ones at closer redshift. In what follows we explain how the original model has been adjusted to account for both factors.

After translating from the spherical system (θ, β, Z) to the Cartesian one $x(\theta, \beta, Z), y(\theta, \beta, Z), z(\beta, Z)$, the noise model at the i-th grid point takes the form $p(g|(x_i, y_i, z_i), C) = \mathcal{N}(\mu_i, C)$. To align the prolonged axis-aligned covariance matrix $\tilde{C} = diag(k \cdot \sigma^2, k \cdot \sigma^2, \sigma^2)$ along the LOS, we employ the corresponding rotation matrix R: $C = R\tilde{C}R^T$.

Given the LOS direction $v = (v_x, v_y, v_z)$ in the Cartesian system, the rotation matrix R can be derived by considering the local frame $u = (u_x, u_y, u_z)$, $s = (s_x, s_y, s_z)$ and v. We impose: $u \perp v$, $s \perp v$ and $u \perp s$. In other words, the dot products $v^T u$, $v^T s$ and $u^T s$ vanish. This leads to an undetermined system. By imposing $u = (0, v_z, -v_y)$ we automatically satisfy $v^T u = 0$. Substituting u in $u^T s = 0$, we obtain

$$v_z s_y - v_y s_z = 0, \quad \frac{v_y s_z}{v_z} = s_y. \tag{3}$$

Using $v^T s = 0$, we get

$$v_x s_x + \frac{v_y^2}{v_z} s_z + v_z s_z = 0, \tag{4}$$

yielding

$$s_x = \frac{-s_z(v_y^2 + v_z^2)}{v_x v_z}. \tag{5}$$

We are left with one free parameter, s_z, that can be assigned arbitrary value (we used $s_z = 1$). After normalization of u, s and v into unit vectors, the rotation matrix is formed as $R = [u, s, v]$ (u, s, v form columns of R).

The model developed so far will not work in the real cosmology since it does not account for the flux limit effect. We are more likely to observe galaxies of the same magnitude close by (at smaller Z) than at high Z. Intuitively, a vote from a galaxy of magnitude M observed at high Z should have higher weight than a vote from a closer galaxy of the same magnitude. Galaxies at large Z are harder to observe than those at smaller Z, and there will be more missing votes from undetected galaxies at large Z. In the probabilistic Hough accumulator (2) each observed galaxy has equal weight $1/N$ when voting for galaxy group positions. A principled treatment of this issue in our model formulation is to replace the weight $1/N$ with a redshift and magnitude specific weight. The modified Hough accumulator thus reads

$$H(x_i, y_j, z_i) = \sum_{q=1}^{N} w(q) \cdot P(i|g_q), \tag{6}$$

where $w(q)$ is the weight given to the q-th galaxy based on its redshift Z_q and absolute magnitude M_q. The weights need to sum to 1 and should be inversely related to the luminosity Schechter function $S(M, Z)$, which for a given absolute magnitude M, gives the density of galaxies of that magnitude at redshift Z [15]:

$$S(M, Z) = \frac{\ln(10)}{2.5} \cdot \phi^* \cdot \left(10^{\frac{M^* - M}{2.5}}\right)^{(\alpha+1)} \cdot \exp\left\{-10^{\frac{M^* - M}{2.5}}\right\}, \tag{7}$$

where $\phi^* = 0.0149 \cdot h^3 \; mpc^{-3}$ is the number density, h is the Hubble parameter, $M^* = -21.35 + 5 \log_{10} h$ is the characteristic magnitude and $\alpha = -1.3$ is the faint-end-slope.

The SDSS survey is complete to an apparent Petrosian magnitude limit of $m \approx 17.77$; however, this can vary somewhat across the sky. Following [3], we adopt a more conservative r-band magnitude limit of $m = 17.5$ to simulate SDSS survey [13]. For each galaxy $q = 1, 2, ..., N$, we estimated its absolute magnitude M_q based on m and the redshift Z_q [15]. We propose the following formulation for the weights $w(q)$ that respects both requirements:

$$w(q) = \frac{S(M_q, Z_q)^{-\gamma}}{\sum_{j=1}^{N} S(M_j, Z_j)^{-\gamma}}. \tag{8}$$

In our experiments, we found $\gamma = 0.3$ to work robustly on the mock data.

4 Experimental Results

We have used the Precision $(TP/(TP + FP))$ versus Recall $(TP/(TP + FN))$ curves in evaluating the group finders, where TP is the number of true positives (correctly detected true groups), FP is the number of false positives (incorrectly detected groups) and FN is the number of false negatives (missed true groups). The precision vs. recall (PvR) curves in Fig. 3 are averages over 10 realizations of background noise in the 2-D data and were obtained by varying the detection threshold Θ. The PHTM method is robust with respect to potentially large amounts of fore/background noise (up to $T = 30$). Note that more direct approaches, such as mixture modeling would end up being swamped with non-group data, even for moderate amounts of background noise (Fig. 2a, b).

For the 3-D realistic data we investigated two settings for the ground truth galaxy groups that needed to be detected: **(+5)** groups containing at least 5 galaxies (including small, harder to detect groups) and **(+10)** larger groups containing at least 10 galaxies.

As an example, Fig. 4a–d shows PvR curves of PHTM on two stripes from two different mock data cones consisting of 34 and 26 galaxy groups respectively. The PHTM is compared with FOF method [5]. Note that while it is very natural to create PvR curves from PHTM (by varying Θ), this turned out to be cumbersome for FOF (modifying free parameters can lead to abrupt changes in performance). Therefore, we report a single value (red star) of (precision, recall) obtained with the parameter setting recommended in [5].

To further compare PHTM with FOF, we identify the closest precision value of PHTM to that of FOF and ask if the corresponding recall by PHTM is similar

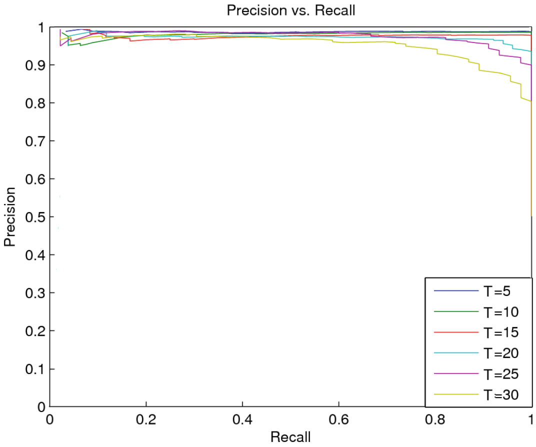
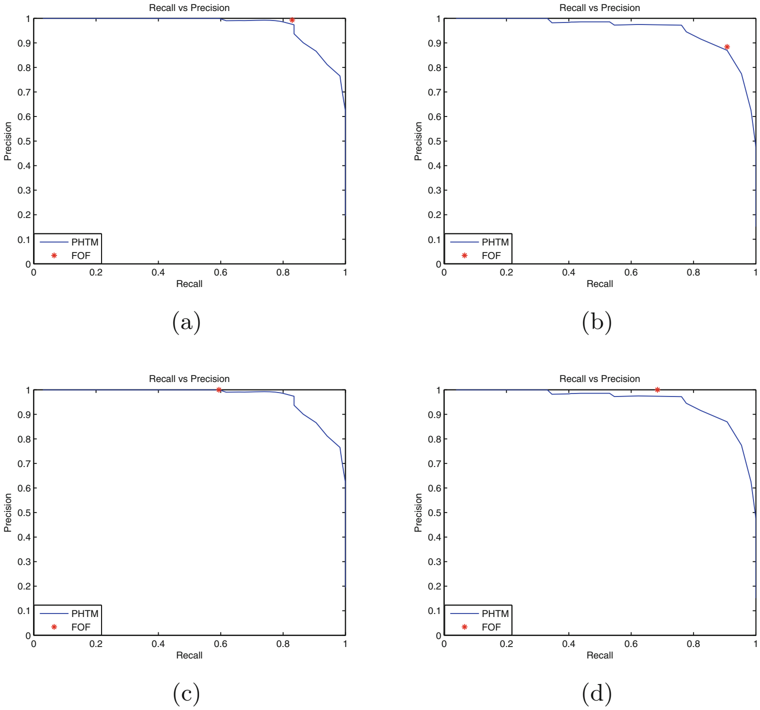


Fig. 3. Average (over 10 realizations of background noise) Precision vs. Recall curves of PHTM on 2D flat mock data



(a)

(b)

(c)

(d)

Fig. 4. Precision vs. recall curves of PHTM on realistic mock data when detecting galaxy groups with: (a) +5 galaxies cone-1; (b) +5 galaxies cone-2; (c) +10 galaxies cone-1; (d) +10 galaxies cone-2

to that of FOF (up to a small tolerance threshold 0.025), or even better beyond the tolerance threshold. For the background intensity $T = 5$, in the groups +5 scenario, out of 20 runs, recall of PHTM was 4 times and 11 times similar and better, compared with FOF. In the +10 group detection, recall of PHTM was 2 times and 15 times similar and better, compared with FOF. Of course, one can repeat the same exercise by fixing the recall to that of FOF and comparing the precision values. In the groups +5 scenario, out of 20 runs, precision of PHTM was 8 times and 5 times similar and better, compared with FOF. In the +10 group detection, precision of PHTM was 12 times and 0 times similar and better, compared with FOF.

For higher background intensity $T = 15$ the figures were as follows: In the groups +5 scenario, out of 20 runs, recall of PHTM was 17 times and 0 times similar and better, compared with FOF. In the +10 group detection, recall of PHTM was once and 18 times similar and better, compared with FOF. Finally, in the groups +5 scenario, precision of PHTM was 4 times and 2 times similar and better, compared with FOF. In the +10 group detection, precision of PHTM was 2 times and 3 times similar and better, compared with FOF.

5 Conclusion

We introduced a novel methodology for a difficult automated detection task - finding galaxy groups embedded in a rich background. The methodology is based on a form of probabilistic Hough transform exploiting a typical signature pattern of galaxy groups known as "fingers-of-God". The model based nature of our methodology enables the user to include prior astrophysical knowledge as an inherent part of the method. The method was first tested in large scale controlled experiments with 2-D patterns and then verified on 3-D realistic mock data (comparing with the well-known friends-of-friends method used in astrophysics. The experiments suggest that our methodology is a promising new candidate for galaxy group finders developed within a machine learning framework.

References

1. Alpaslan, M., et al.: Galaxy and mass assembly (gama): the large scale structure of galaxies and comparison to mock universes. MNRAS **438**(1), 177–194 (2014)
2. Astone, P., Colla, A., D'Antonio, S., Frasca, S., Palomba, C.: Method for all-sky searches of continuous gravitational wave signals using the frequency-hough transform. Phys. Rev. D **90**, 042002 (2014)
3. Berlind, A.A., Frieman, J.A., Weinberg, D.H., Blanton, M.R., Warren, M.S., Abazajian, K., Scranton, R., Hogg, D.W., Scoccimarro, R., Bahcall, N.A., Brinkmann, J., Gott, J., Richard, I., Kleinman, S., Krzesinski, J., Lee, B.C., Miller, C.J., Nitta, A., Schneider, D.P., Tucker, D.L., Zehavi, I.: Percolation galaxy groups and clusters in the sdss redshift survey: identification, catalogs, and the multiplicity function. A.J.S **167**, 1–25 (2006)
4. Duarte, M., Mamon, G.A.: Maggie: models and algorithms for galaxy groups, interlopers and environment (2014). arXiv:1412.3364

5. Eke, V.R., et al.: Galaxy groups in the 2dFGRS: the group - finding algorithm and the 2PIGG catalog. MNRAS **348**, 866 (2004)
6. Hollitt, C., Johnston-Hollitt, M.: Feature detection in radio astronomy using the circle hough transform. PASA **29**, 309–317 (2012)
7. Huchra, J.P., Geller, M.J.: Groups of galaxies. I - nearby groups. APJ **257**, 423–437 (1982)
8. Liu, H.B., Hsieh, B., Ho, P.T., Lin, L., Yan, R.: A new galaxy group finding algorithm: probability friends-of-friends. APJ **681**(2), 1046 (2008)
9. Llebaria, A., Lamy, P.: Time domain analysis of solar coronal structures through hough transform techniques. In: Mehringer, D., Plante, R., Roberts, D. (eds.) Astronomical Data Analysis Software and Systems VIII. Astronomical Society of the Pacific Conference Series, vol. 172, p. 46 (1999)
10. Mukhopadhyay, P., Chaudhuri, B.B.: A survey of hough transform. Pattern Recogn. **48**(3), 993–1010 (2015)
11. Mushotzky, R.: Clusters of galaxies: an x-ray perspective. Clusters of Galaxies: Probes of Cosmological Structure and Galaxy Evolution, p. 123 (2004)
12. Navarro, J., Frenk, C.S., White, S.: The structure of cold dark matter halos. APJ **462**, 563 (1996)
13. Pearson, R.J., Ponman, T.J., Norberg, P., Robotham, A.S.G., Farr, W.M.: On optical mass estimation methods for galaxy groups. MNRAS **449**(3), 3082–3106 (2015)
14. Ramella, M., Geller, M.J., Huchra, J.P.: Groups of galaxies in the center for astrophysics redshift survey. APJ **344**, 57–74 (1989)
15. Schechter, P.: An analytic expression for the luminosity function for galaxies. APJ **203**, 297–306 (1976)
16. Storkey, A.J., Hambly, N.C., Williams, C.K.I., Mann, R.G.: Cleaning sky survey data bases using hough transform and renewal string approaches. MNRAS **347**(1), 36–51 (2004)
17. Tino, P., Zhao, H., Yan, H.: Searching for coexpressed genes in three-color cdna microarray data using a probabilistic model-based hough transform. IEEE/ACM Trans. Comput. Biol. Bioinform. **8**(4), 1093–1107 (2011)
18. Tyson, J.A., Valdes, F., Jarvis, J.F., Mills, A.P.: Galaxy mass-distribution from gravitational light deflection. APJ **281**(2), L59–L62 (1984)