agronomy

MDPI

Article

Unsupervised Greenhouse Tomato Plant Segmentation Based on Self-Adaptive Iterative Latent Dirichlet Allocation from Surveillance Camera

Qifan Cao and Lihong Xu *

College of Electronic and Information Engineering, Tongji University, Cao'an Road, NO.4800, Shanghai 201804, China; caoqifan@tongji.edu.cn

* Correspondence: xulihong@tongji.edu.cn; Tel.: +86-136-363-62684

Received: 16 January 2019; Accepted: 1 February 2019; Published: 16 February 2019

Abstract: It has long been a great concern in deep learning that we lack massive data for high-precision training sets, especially in the agriculture field. Plants in images captured in greenhouses, from a distance or up close, not only have various morphological structures but also can have a busy background, leading to huge challenges in labeling and segmentation. This article proposes an unsupervised statistical algorithm SAI-LDA (self-adaptive iterative latent Dirichlet allocation) to segment greenhouse tomato images from a field surveillance camera automatically, borrowing the language model LDA. Hierarchical wavelet features with an overlapping grid word document design and a modified density-based method quick-shift are adopted, respectively, according to different kinds of images, which are classified by specific proportions between fruits, leaves, and the background. We also utilize the feature correlation between several layers of the image to make further optimization through three rounds of iteration of LDA, with updated documents to achieve finer segmentation. Experiment results show that our method can automatically label the organs of the greenhouse plant under complex circumstances, fast and precisely, overcoming the difficulty of inferior real-time image quality caused by a surveillance camera, and thus obtain large amounts of valuable training sets.

Keywords: greenhouse tomato plant; image training set; unsupervised image segmentation; latent Dirichlet allocation; surveillance camera; self-adaptive word document assignment; density-based estimation

1. Introduction

The vision information of greenhouse crops is of great significance for plant phenotype analysis, which has been applied to provide guidance to improve crop cultivation in many studies [1–5]. It is obvious that an automatic, accurate, and high-throughput imaging processing technique—with high sensitivity for plant phenotypic research—not only lays a visual foundation for analyzing the effect on the environment in production, but is also conducive to comprehensive research of internal and external factors on the physical and biochemical characteristics of plants. Besides, it serves as a non-destructive analysis method to make contributions to directive breeding, crop identification, and yield estimation. In this way, the growth of plants and the optimal regulation of intelligent greenhouse control system can be well monitored and strengthened [6–9].

As there has been an increasing interest in deep learning, we also see wide application of 2D images in the field, such as plant classification, organ detection, maturity estimation, pest disease examination, and so on in the literature. However, there do exist many undesirable problems. On the one hand, we lack large amounts of reliable training data based on the real greenhouse environment in further supervised learning process. That is to say, there is little open source data available in relation

to greenhouse plants, thus we have to depend on inefficient artificial means to obtain images and label them one by one, which will usually be poorly-segmented and not cost-effective. On the other hand, different from crop images in open-air fields, pipes, glass, and other greenhouse mechanical equipment are also regarded as background interferences. Note that not only do densely-packed branches, overlaps between fruits, and reflection deriving from supplementary lighting increase the difficulty in acquiring plant information, but the unsatisfactory image quality caused by greenhouse surveillance cameras also act as an adverse effect in image processing, as compared to pictures taken by human. From the perspective of quantity and quality of the image of the greenhouse crop, it is of great value to conduct research about how to segment and label such images from the camera accurately and rapidly, in order to provide abundant data for further phenotypic study.

Until now, a diverse range of image segmentation algorithms have been presented in the agriculture field, among which there are color-based threshold methods [10,11], active contour methods combined with prior knowledge [12,13], neural network-based methods [14], and graph- clustering methods [15–17]. In order to obtain more accurate results, supervised algorithms [18,19] are also involved. We notice that some statistical approaches surpass many other methods and have been gaining more and more attention. These approaches are often applied by modeling the image to achieve the maximum probability distribution of each pixel, or by solving an optimization problem through minimizing an energy function [20–23].

In recent years, an unsupervised generative probabilistic model LDA (latent Dirichlet allocation) has come into people's view, which is a statistical method to deal with natural language processing problems under varied situations. The LDA method is based on a word topic model, which considers documents as random mixtures over latent topics, and represents each topic as a multinomial probability distribution over orderless words. Objects can be labeled by calculating the topic probability of words in the documents using learning approaches, such as the variational method or Gibbs sampling in a statistical way. Later, it became common, with widespread use, to tackle problems in image processing and this developed rapidly [24,25], which takes advantage of various encoding words to achieve a better result of image segmentation or labeling.

A spatial latent Dirichlet allocation (SLDA) topic model [26] encoded the spatial structure of visual words as a random hidden variable, instead of depending on the partition of a word document design as a prerequisite in LDA's generative procedure. Niu et al. [27] extended a supervised LDA model (DiscLDA) [28], and fused a location feature into it to form a spatial-DiscLDA (S-DiscLDA), in order to mitigate the inefficiency of the spatial structure resulting from orderless visual words. A spatially-coherent latent topic model (Spatial-LTM) [29] enforces the spatial coherency of the model and provides a unified representation in a hierarchical way, by over-segmented image regions of homogeneous appearances, which has the ability to segment and classify objects with multiple instances. Li Z et al. [30] proposed a model bag-of-topics (BoT) to calculate the similarity between the images, instead of bag-of-words (BoW), for discovering the abstract 'topics' from the words, both reducing the dimension of the image representation and improving the retrieval performance. Niu Z et al. [31] also presented an LDA model, called Dirichlet trees (LDA-MDT), which incorporates multiple 'must-links' as prior knowledge into topic modeling, improving topic coherence for object discovery and localization.

In this paper, we elaborately analyze the complexity of greenhouse tomato plant images and the undesirable picture quality brought by a fixed camera. Given that the image structure varies a lot in the photos from different shooting distance, we put an emphasis on codebook construction as well as document assignment strategy through appropriate spatial structure encoding of LDA. Then we propose the SAI-LDA algorithm to automatically segment and label the plant according to the image type, which is determined by camera focal length. This modified LDA algorithm is modeled with the features of the pixels, which are clustered into some classes on the basis of the maximum probability, and achieve its best performance by further iteration. Experiments show excellent results to prove that this algorithm outperforms the traditional LDA and state-of-the-art segmentation methods in the

agriculture field, getting over the challenge of poor-quality source data from cameras and providing labeled images with high accuracy. Thus, we deal with this tough problem in a fully unsupervised setting so that more reliable label information, as compared to manual labeling, is obtained and a sound training set can be available.

2. Materials and Methods

2.1. LDA (Latent Dirichlet Allocation)

The latent Dirichlet allocation, LDA, first proposed by Blei, Ng et al. [32], is one of the most popular topic generation models, which contains a three-level structure of word, topic, and document. Recently, it has developed rapidly and become distinguished in the field of image processing such as image recognition, classification, annotation, and so on [33,34]. Its fundamental conception, bag-of-words model (BoW) [35,36], was originally used in distinguishing hidden information in a large collection of corpus [37,38] and conversing the information of the pixels to non-ordered visual words. As an unsupervised generative probabilistic model, its documents are viewed as a mixture of topics, sharing a common Dirichlet priori. Similarly, each topic is characterized by a probabilistic distribution over the vocabulary of words in the collection as well. In other words, documents are generated by first drawing topic proportions while topics are discovered by finding the appropriate sets of latent variables, given the topic proportions of documents. Actual words are sampled and clustered into K topics by drawing from the multinomial distribution over the codebook based on a frequency histogram. Words which correspond with their distributed topic and with the highest probability in each topic usually stand for what the topic is according to LDA. Figure 1 shows the LDA Bayesian network structure.

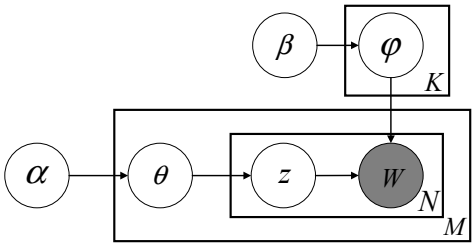

Figure 1. LDA Bayesian network structure.

M denotes the documents; N represents the set of words in M; K is the number of topics; W represents the sampled words from the document; z represents the topic for the denoted word; θ and φ are multinomial parameters sampled from the topic distribution and the word distribution, respectively. α and β denote the prior parameters set for the document topic distribution and topic word distribution, respectively. They determine the sparsity or the uniformity in the collection of documents. The higher α is, the more topics a document will be assigned. The higher β is, the more words a topic will contain. The generative procedure is shown in Figure 2 in detail.

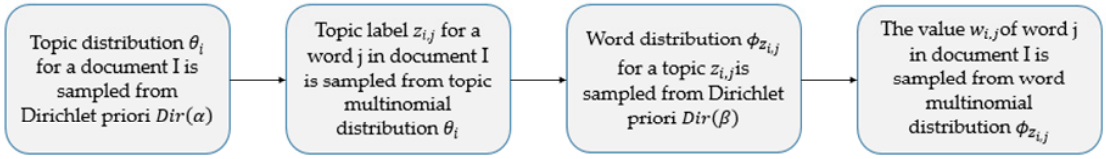

Figure 2. The generative procedure of LDA.

The probability of observed data is computed and then maximized to form the joint distribution so that the latent variables, α and β can be inferred as

$$p(w_i, z_i, \theta_i, \Phi|\alpha, \beta) = \prod_{j=1}^{N} p(\theta_i|\alpha)p(z_{i,j}|\theta_i)p(\Phi|\beta)p(w_{i,j}|\theta_{i,j}) \quad (1)$$

After integrating θ_i, Φ, and summing z_i, the maximum likelihood estimated for the word distribution oriented at one document can be obtained as

$$p(w_i|\alpha,\beta) = \int_{\theta_i}\int_{\Phi}\sum_{z_i} p(w_i, z_i, \theta_i, \Phi|\alpha,\beta) \quad (2)$$

According to the maximum likelihood estimation of $p(w_i|\alpha,\beta)$, the parameters in the model can finally be estimated by means of Gibbs sampling.

2.2. Limitation Analysis of LDA and Practical Situation

However, challenges do exist when utilizing LDA to solve image segmentation problems, that is, visual feature definition, limitation of the BoW model and the assumption of visual problem. As we know, it is of great significance to extract accurate visual features for each pixel in the image. We must have a clear understanding of how to choose appropriate features, how to cluster them into discrete words, and finally quantify these visual features according to the vocabulary (codebook), so that each pixel is represented as the corresponding visual word. Only when the visual feature extraction rule is confirmed can the corresponding theme be generated. When it comes to the BoW model and visual assumption, how to design the topic generation model that can contain thorough space information is definitely pivotal. The traditional LDA method maps the whole image to a single document or utilizes non-overlapped rectangular regions as documents. These two ways are in accordance with the assumption that if visual words belong to the identical class, they are likely to co-occur in the same image or in close proximity to each other. Nevertheless, it is a pity that these simple document allocations only depict poor local region information under the circumstance that parts of the same object are in different but neighboring documents, resulting in misclassification between two classes. Figure 3 shows the LDA algorithm flow for image segmentation.

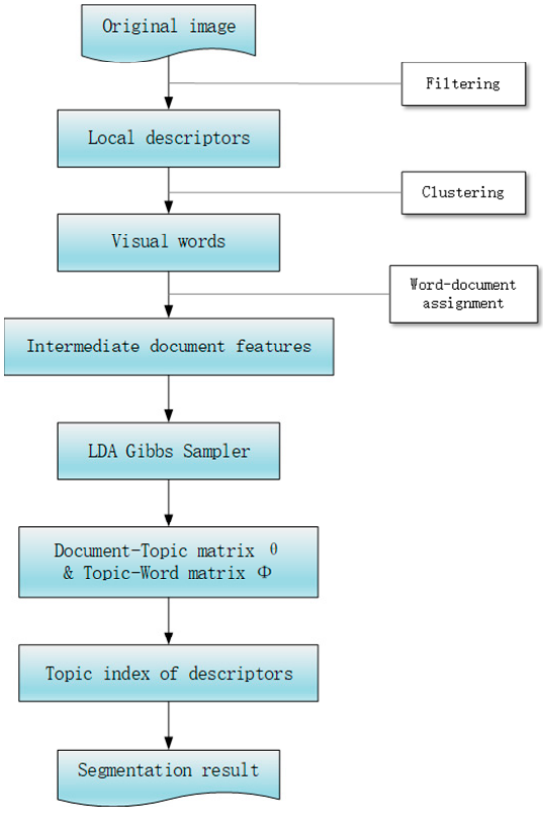

Figure 3. LDA segmentation algorithm flow.

To sum up, those rough document allocations, which may be feasible for some natural language processing in the BoW model, will lead to deficiency in detection of co-occurrence information between images, so that a huge gap between low-level visual words and high-level semantics information is generated. Figure 4 is the elementary process of word document distribution, which clarifies the operation of the traditional LDA and its encoding weakness.

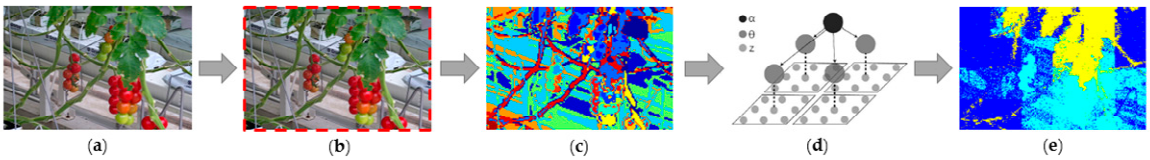

Figure 4. Process of word document encoding based on fixed non-overlapping grid. (**a**) Original image; (**b**) Consider the whole image as one document; (**c**) Quantified visual words; (**d**) Word document model: α is the prior parameter for the document topic distribution; θ is the multinomial parameter for the topic distribution; z is the corresponding topic for the denoted word; (**e**) Segmentation result.

With the aim of acquiring quantities of greenhouse tomato images, it is unrealistic to simply rely on manual photography which takes too much time and energy. In this regard, we need to resort to the camera installed inside the greenhouse to take photos automatically and quickly. Unfortunately, many unfavorable factors affect the image quality, such as single shooting angle, occlusion of fruit or leaves, out-of-focus targets, edge distortion, and so on. Furthermore, the distance between the target plant and the lens is different, which results in distinct location and proportion of the region of interest in the picture. If the picture is taken from a remote view, tomatoes usually appear in clusters. Besides, the background occupies the majority of the picture while the plant occupies the small remaining part. Conversely, when the camera zooms in, the plant makes up the most part. As a result, it is quite a hard work to grasp the co-occurrence information of the plant images due to the complexity and manifold types of the pictures. In case the feature and document are not adaptively selected and allocated according to the image type, the words of a document will be inaccurate to express a specific topic, and the object will also lose its crucial region or edge information.

In a word, the traditional LDA method will meet bottleneck in grasping the spatial structure information under the background of practical greenhouse application. If the local feature description and global structure summarization can be well designed, the LDA algorithm will be enhanced a great deal. So here arises our priority that how to adaptively optimize the LDA algorithm based on the fact that different types of tomato plant images exist in our database, according to the two given considerations above.

2.3. Self-Adaptive Latent Dirichlet Allocation(SA-LDA)

2.3.1. Self-Adaptive Image Type Judgement

First of all, we make an analyzation of the 20 acquired pictures from our dataset (the collection process can be seen in Section 3.1), half of which are close shots while the rest are distant shots. Intuitively speaking, the background of distant shots occupies most parts of the picture, including the ground and nutrient tanks, and the proportion of leaves and fruits is relatively less. Especially for fruits, there are clusters of tomatoes but the size of them are quite small. For the close shots, the proportion of leaves is larger due to the shortening of focal length. Meanwhile, the number of fruits decreases while the proportion scales up with little background left, which only exists as the gap between leaves. Therefore, we can judge these three parts through their proportion of each other.

Since there is no need to make accurate segmentation for three parts at this stage, we utilize color quantification to roughly estimate the image structure. For the leaf part, we use the excess green index and modify coefficients by (1.3R-G-B) according to the actual situation of our greenhouse. For the fruit part, we have studied several color spaces such as (R-G), (R-B), (R-G-B) and other combinatorial

operators as feature quantities to perform algebraic operations. We found that the saliency of tomato fruits with the subtraction of red and blue components by (R-1.5B) was significantly different from those of leaves and the background. The fruits are bright and the rest of the picture is dark so that we can distinguish each part well. Thus, the gray value (R-1.5B) is used to extract the fruits, as shown in Figure 5. Figuring out the part of fruits and leaves, we reckon that the rest of the image is background. In this way we can transform the original three-dimensional problem into one-dimensional problem and classify images by simple color feature analysis.

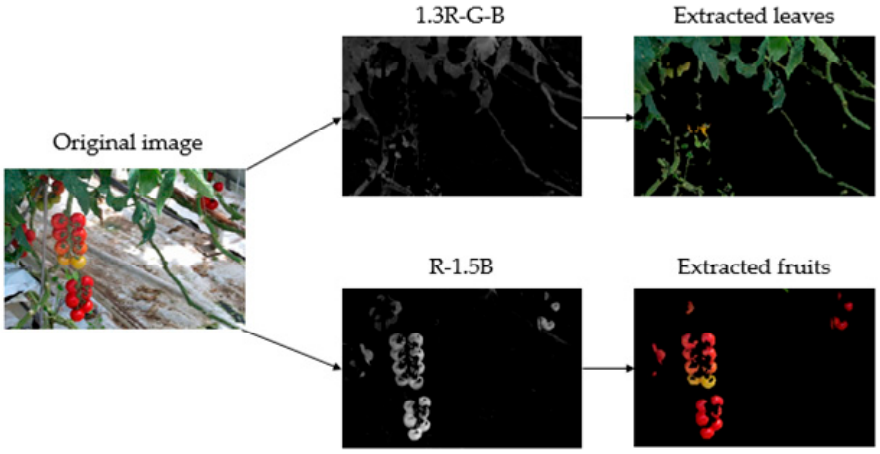

Figure 5. Extraction of tomato organs: algebraic operations in RGB color space.

We then calculate the proportion of the three parts in 20 pictures, seen from Figure 6. The former 10 are distant shots while the latter 10 are close ones. We work out that the average proportion of expected fruit, leaf and background for distant shots are 8.54%, 41.69%, 44.17%, respectively, while 23.04%, 66.38%, 13.75% for close shots. Data shows the proportion of fruit and background changes significantly.

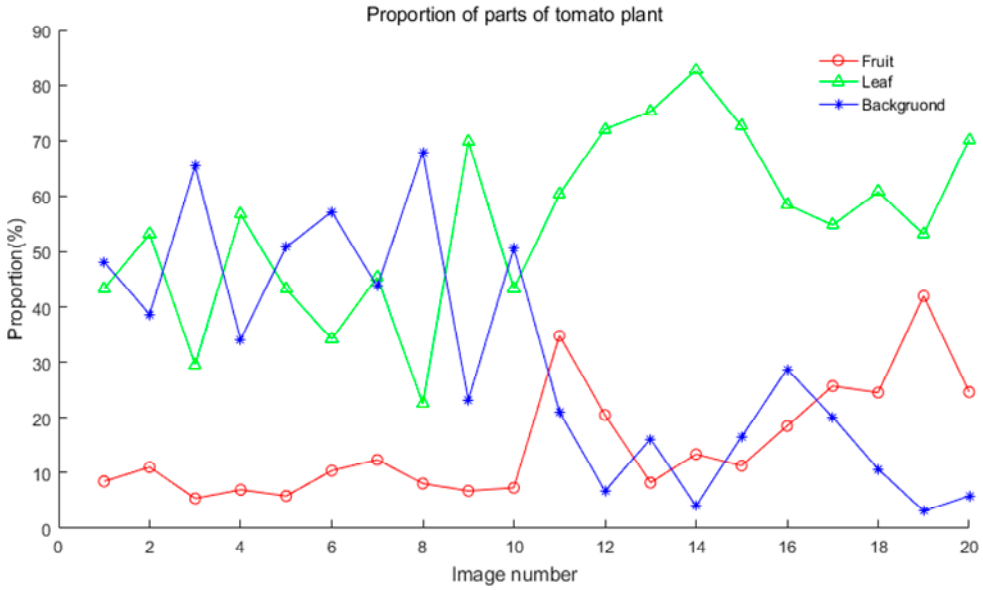

Figure 6. Proportion of tomato organs.

In a word, if there exists larger proportion of fruits and leaves while the ratio of the background is smaller, the image is more likely to be classified into close shots. In this regard, we define a distance determination formula (*DD*)

$$DD = \alpha \cdot \sqrt[3]{\mathrm{P_l}} + \beta \cdot \log_2 \mathrm{P_f} + \gamma \cdot \mathrm{e}^{-0.4 \times \mathrm{P_b}} \quad (3)$$

P_l, P_f, P_b stand for the proportion of leaves, fruits and background. α, β and γ are weight parameters intended for each class. As we discussed above, the proportion of fruits or background varies greatly between the two kinds of images, so we pay more attention on them and define α = 0.2, β = 0.55, γ = 0.35. We calculate the score of each image in Figure 7.

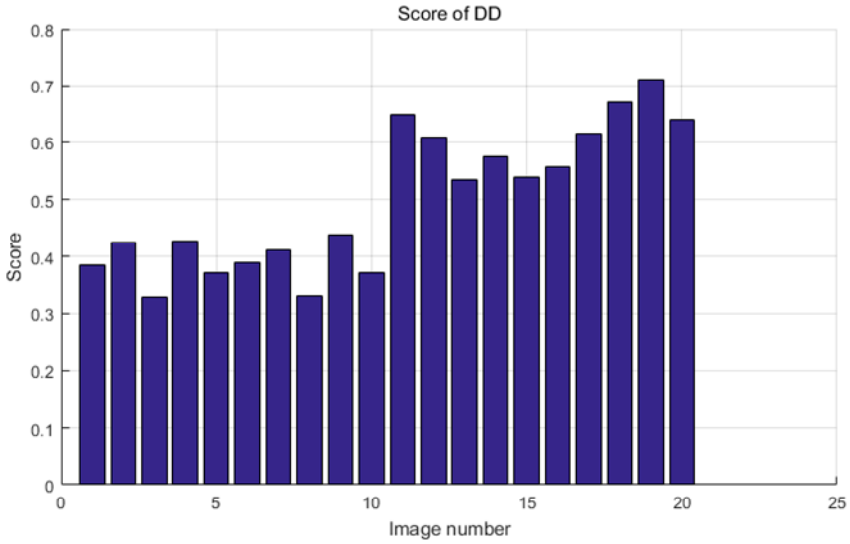

Figure 7. Distance determination (DD score).

Through the bar graph above, the average DD score for distant image and close shots are 0.3882 and 0.6117, so we choose the threshold value 0.5 as the critical point of judgment. When we acquire the proportion of the three parts of the picture according to the color feature, we substitute them into the DD formula and calculate the score. If DD $\leq$ 0.5, we confirm that the picture belongs to the distant one while when DD > 0.5, we assume the picture a close shot.

2.3.2. Wavelet-Based Feature Construction

The design of visual words in the BoW model sets as a preliminary and important part in feature encoding to intrinsically approximate the discriminative structure of images. As we discussed above, from a long-distance perspective, the image is under cluttered background and contains blurred boundary information between the objects. Especially under the circumstance that immature fruits and green leaves are the principal part, it usually leads to great confusion between fruits, leaves, and other disturbing factors. To this regard, we cannot simply use color feature to describe the local pixels, but should also focus more on the description of texture features in the stage of visual word construction. However, we should pay more attention on how detailed the word expresses. The more detailed the feature is, the higher dimension grows, which sometimes hinders the capability of the algorithm to extract the topic and results in great difficulty to achieve convergence state.

It is known that the discrete wavelet transform (DWT) is utilized as a popular multiresolution technique to obtain hierarchical feature information and is efficiency effective in computation through processing at each scale [39]. On each level, the low-frequency wavelet coefficients represent the rough outline of the overall image while the high-frequency ones are viewed as the portrayal of the edge strength. With the altering of resolution, the texture of the image will change or even disappear. Based on these characteristics, we discover more on a finer scale to complement for the feature loss on a

coarser scale. Time and frequency locality as well as multi-resolution make it possible for wavelet transform to process all signals from high-frequency domain to low-frequency domain, which means information from boundary to internal region. Since wavelet transform makes features of each pixel well expressed, when the proportion of tomato plants in the picture is small, the disparity between plants and the background can be exactly detected at each scale. Besides, high compression ratio enables wavelet transform to represent complex signals with a small number of wavelet and scale coefficients, with which it reduces the feature dimension and the computation. Thus, we can make use of these coefficients to merge feature into visual words, facilitating the later codebook construction of LDA.

The image is decomposed into four sub-bands, LLi, LHi, HLi, and HHi (i = 1, 2, 3 and it stands for the image level) at each scale as Figure 8 shows. LL is a low-frequency sub-band, which serves as a smooth approximation of the original image. The other three stand for high-frequency sub-bands, among which LH, HL, and HH are the detail images along horizontal, vertical, and diagonal directions, respectively. In most papers, we found the LL sub-band is commonly used as the texture approximation of the original image for calculation in the next level. However, they do ignore some dominant information left in the rest of the sub-bands.

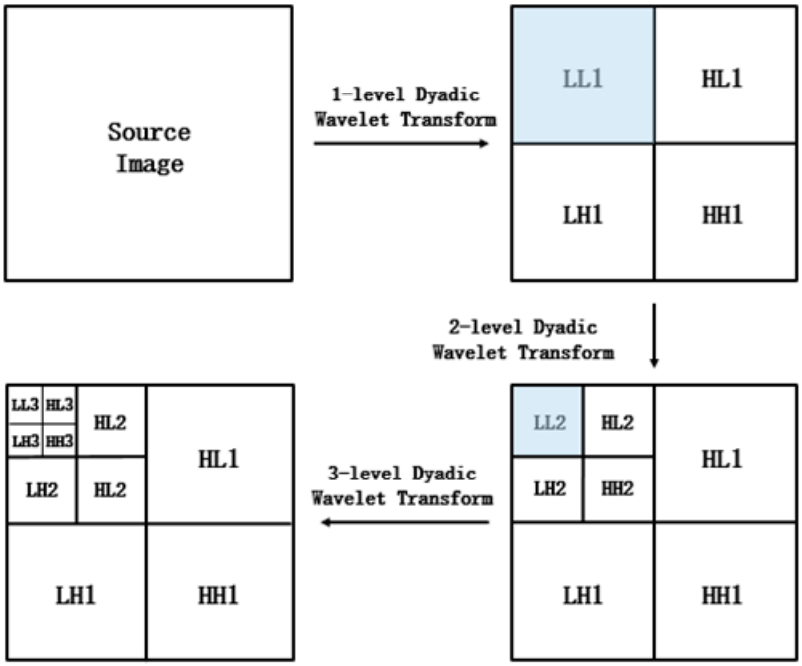

Figure 8. Sub-bands separated by a three-level dyadic discrete wavelet transform (DWT).

Therefore, we introduce discreet wavelet packet transform (DWPT) [40] to decompose every sub-band further which contains the approximate description in LL as well as detailed ones in LH and HL at each level. We exclude the HH sub-band with the cognition that it contains majority of the noise so that it brings about worse performance. Hence in this way the wavelet transform is capable of providing the detailed analysis of the image data and extract more features in varied directions. Instead of Daubechies-4 wavelet filters used in [40], we modify the wavelet base function and make it further decompose at every sub-band with symlet wavelet filters which gives 4, 16 and 64 sub-bands at each level respectively. The support range of symlet wavelet is 2N-1, where N is the vanishing moment and we define N = 3. So the total sub-bands are equal to fifty-two ($4 \times (1 + 3 + 9) = 52$) from the original eighty-four sub-bands ($4 \times (1 + 4 + 16) = 84$). It has a good regularity and symmetry compared to Daubechies filters, which is consistent in continuity, support length and filter length. In other words, it is smoother and has the capacity of reducing the phase distortion in signal analysis reconstruction to a certain extent. Afterwards we reduce the 52 features by taking averages of standard

deviations and energies of sub-bands at each of the three decomposition levels. The standard deviation (SK) and energy (EK) are first computed using Equation (4) and Equation (5) respectively.

$$\mathrm{SK} = \sqrt{\frac{1}{MN}\sum_{i=1}^{M}\sum_{j=1}^{N}(B_k(i,j)-\mu_k)^2} \tag{4}$$

$$\mathrm{EK} = \frac{1}{MN}\sum_{i=1}^{M}\sum_{j=1}^{N}|B_k(i,j)| \tag{5}$$

B_k is kth decomposed sub-band, MN, μ_k are the size and mean of kth decomposed sub-band respectively. Since we hope to extract the most information with the least feature vector considering the practical computational expense, we then figure up the average energy and average standard deviation of all sub-bands at each scale with Equation (6) and Equation (7) to further reduce the dimension.

$$S_{ka} = \frac{1}{L}\sum_{i=1}^{L} S_{ki} \tag{6}$$

$$E_{ka} = \frac{1}{L}\sum_{i=1}^{L} E_{ki} \tag{7}$$

k is the decomposition level and L is the number of kth sub-bands without HH sub-band. S_{ki} and E_{ki} represent the average of standard deviations and energies at kth level. So, the texture feature vector (TFV) has only six elements left as

$$TFV = \{S_{a1}, S_{a2}, S_{a3}, E_{a1}, E_{a2}, E_{a3}\}$$

Furthermore, we convert the image into CIE *Lab* color space and extract two channels L and a as color information. All in all, our feature vector consists of eight elements considering both color and texture, which makes good use of wavelet characteristics and mine the image information as much as possible at multiple scales. So, feature vector (FV) is

$$FV = \{S_{a1}, S_{a2}, S_{a3}, E_{a1}, E_{a2}, E_{a3}, L, a\}$$

When it comes to the document assignment, instead of taking the whole image as one document, which neglects spatial relationship between different classes of objects, we choose many overlapped grip patches as documents. Patches are orderly arranged with an interval a (a = 25 pixel) as shown in Figure 9, which corresponds to a better visual assumption. This method considers both the tolerance to feature differences and the case that one object is dispatched to different documents close in space. In this way, single part in several patches can increase the probability to be allocated to the right document through joint sampling. After we densely sample the modified DWPT descriptors on each patch, a popular clustering method k-means is employed so that N cluster centers are obtained which makes up of the codebook of size W. We find the corresponding cluster center of each descriptor by choosing the minimum Euclidean distance and finally quantify them into visual words according to the codebook. We simplify the prior steps below:

- Apply three-level discrete wavelet packet transform decomposition and obtain 52 sub-bands abandoning HH ones. Compute standard deviations and energies of each level using Equations (4)–(7) to signify the texture feature;
- Extract channel L and a from CIE *Lab* color space for color feature;
- Input eight-dimension integrated feature vector for words{w}, overlapped regions for documents{d} and number of class K into LDA model. Compute the topic-word parameter matrix by Gibbs sampling and classify the image to K parts according to the probability.

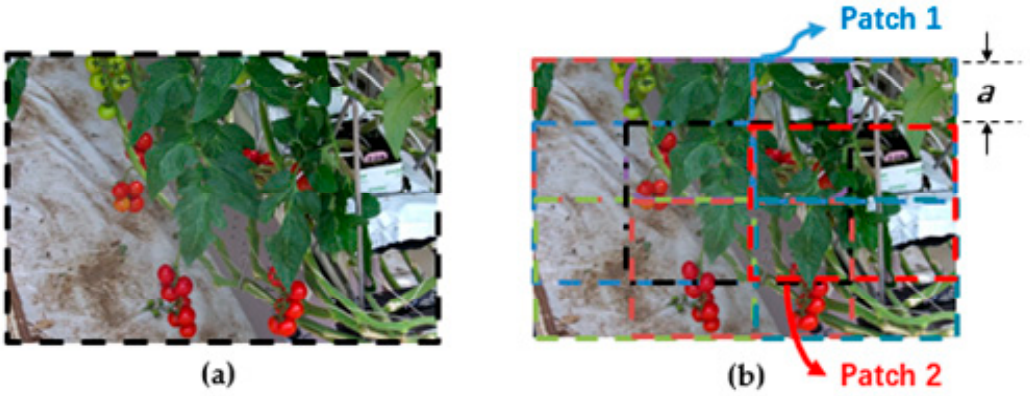

Figure 9. Word document encoding based on an improved document partition. (**a**) The traditional LDA document construction regards the whole image as a single document; (**b**) Overlapping rectangular patches are adopted with interval a to construct documents, where a = 25 pixel.

2.3.3. Density-Based Document Allocation

Note that the difference between shooting distance will bring about huge gap within the same method. The tomato fruits and leaves in images discussed above are relatively small in size so that rectangular overlapping documents are suitable with multi-resolution wavelet feature extraction. However, this document assignment has less tolerance to different parts of the object with different characteristics to some extent. As for close shots, firstly, the plants often reside in the center of the image, and the number of pixels is large. The rigid structure of rectangular documents will have poor performance in expressing the image characteristics clearly due to the fact that the number of visual words in each document are the same. Meanwhile, the unnatural rectangular area limits the ability of documents to contain more accurate information depending on practical situation. Secondly, the design of rectangular documents can only focus on pixels within a certain distance. While for those words far away from each other, it seems to have difficulty in clustering them into one document, which implies that the lack of spatial information coordination will probably lead to poor topic allocation. Futhermore, the proportion of bright spots or shadows caused by illumination is larger than distant shots in the whole image. As a result, we not only need to redesign the document for a better spatial structure, but also deal with the impact of various illumination effect first.

In this paper, we introduce illumination correction of homomorphic filtering as a preliminary process to weaken the negative influence. It is a kind of non-linear filter and belongs to the frequency domain processing method based on the principle of illumination reflection, of which the function is to eliminate the problem of uneven lighting by adjusting the gray range of the image. Due to the contrast enhancement and brightness compression in frequency domain, it can avoid illumination changes and sharpen edge details to remove noise. The image can be viewed as the product of the radiation component and the reflection component. After Fourier transform, we design a frequency domain filter $H(u, v)$ and get results in the spatial frequency domain after Fourier inversion. The major formulas are

$$H(u, v) = (\gamma_H - \gamma_L) H_{\mathrm{hp}}(u, v) + \gamma_l \tag{8}$$

$$H_{hp} = 1 - \exp\left[-c\left(D^2(u, v)/D_0^2\right)\right] \tag{9}$$

γ_H and γ_L control the range of filter amplitude while H_{hp} is Gaussian high-pass filter. c is a constant that controls the shape of the filter, that is the slope of the transition from low frequency to high frequency.D_0 corresponds to the value when $H(u, v)$ approaches γ_H. Here we set γ_H = 2, γ_L = 0.5, c = 1.5, D_0 = 100. Figure 10 shows the differences in histogram after homomorphism process.

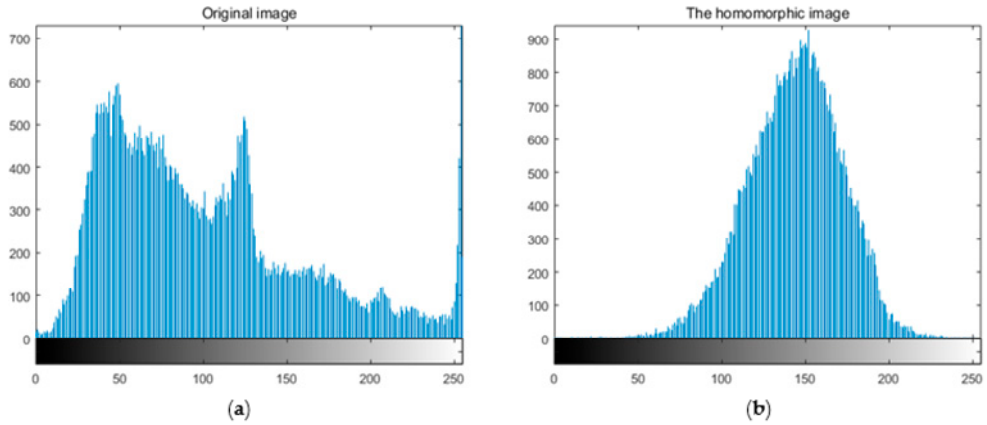

Figure 10. (**a**) Histogram of the original image; (**b**) histogram after the homomorphic filtering.

After executing homomorphic filtering process, we have realized the compression of dynamic range and the enhancement of image contrast, eliminating the influence of non-uniform illumination without losing image details to a great extent. Now we focus on how to design the document. It is known that the distribution of image data is flexible and has no fixed pattern, so nonparametric methods should be used here to conduct the unknown data-generating density estimation and the encoding of image structure. Different from the original strategy with rectangular documents, what we consider is to make use of local maximum points found by probability density function of the image as key points. Then we define clusters as basins of attraction points and form documents of LDA. Here quick shift is utilized to determine which document a word belongs to. Briefly speaking, the documents are represented by accumulated modes, thus clusters cores with high similarity come into being, which can be regarded as a more effective spatial structure to facilitate classification ability of LDA.

As for quick shift, it is an iterative mode seeking algorithm that identifies the densest location in a certain search space of data points [41]. Prior to this method, mean shift [42] and medoid shift [43] have been brought out and widely used in data clustering. Each point x_i is moved towards a mode of P(x) along the trajectory $y_i(t)$ uphill following the gradient $\nabla \mathrm{P}(y_i(t))$. At last, all the points will converge to the same mode and a cluster is formed. Figure 11 below shows the different way of moving path between those three algorithms. Recently, many extensions and improvements of such mode-based methods have been developed and proposed in the statistics literature [44–46].

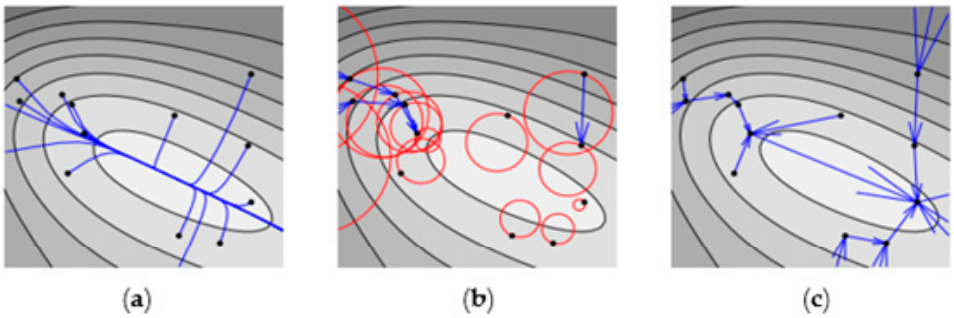

Figure 11. Mode seeking algorithm. (**a**) Mean shift: modes converge by estimating probability density kernels; (**b**) medoid shift: local modal approximation based on sample point-neighborhood weighted estimation; (**c**) quick shift: directly connect the sampling point to the highest density one.

Traditionally speaking, quick shift, which acts as a faster alternative to the well-known method, mean shift, proceeds by connecting all the sample points into a single tree. The root of the tree is the pixel which corresponds to the highest density mode in the image. Then it repeatedly moves each one to its closest sample point that has a higher empirical density by breaking the branches of the tree that

are longer than a threshold τ. No need for any iterative calculation, samples that belong to the same points of convergence are in the same cluster and are taken to be the estimated recovery of the modes.

Although traditional quick shift does surpass mean shift both in estimated result and time complexity; however, we also find some drawbacks in this sample-based method. One is for its poor representation of clusters when images have arbitrary shape of high-density regions, which makes it easy to detect excessive apparent modes, thus over-segmentation occurs. The other is its common solution to combat over-segmentation by increasing the kernel bandwidth, which probably give rise to serious deviation from the intended density. Both of these shortcomings mean an unsatisfied image estimation, which will definitely weaken the rationality of the document design.

In this regard, we make some modification of the density-based method, the main idea of which is to identify clusters on locally high density rather than singleton point-modes as final destination [47], seen from Figure 12. We first connect the detected modes and form regions, estimating the center of these regions as modal-sets, in other words, cluster-cores. Afterwards, we run quick shift based hill climbing procedure on each remaining sample until it reaches a cluster-core, which can represent high-confidence regions with the anti-interference performance. Besides, the extra parameter τ is no longer needed in our algorithm because we set our stopping condition on reaching the cluster-core. The character can avoid the hard determination of τ in practice as there is hardly such an appropriate setting which works everywhere in the input space. Moreover, it can still trade off under-and-over-segmentation of the modes in a straightforward manner and recover modal-sets at varying density levels.

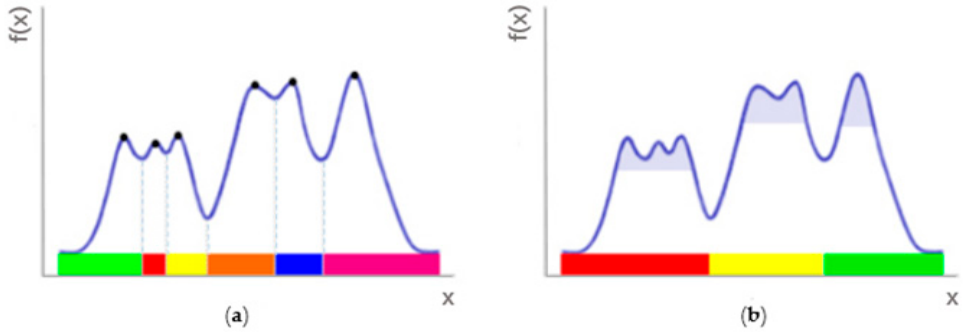

Figure 12. (**a**) High-density regions may have variations and too many apparent modes appear, leading to over-segmentation; (**b**) using modal-sets (shaded) not only allows disturbances within limits of different shape of regions of high density, but also matches various density levels, regarding them as the same cluster.

The pixels in an image include coordinate space in spatial domain, $x^s = [px, py]$ and color space CIE LUV in range domain, $x^r = [l, u, v]$. We first initialize the joint feature space of the image data by transforming each pixel into the five-dimensional vector $[px, py, l, u, v]$. Our segmentation is done by clustering this five-dimensional dataset. We have samples $X_{\{n\}} = \{X_1, X_2, \ldots X_n\}$ drawn from the image distribution with density f on R^d. For each x in the support of f, we set $r_k(x) = \min\left\{r > 0 : \left|B(x,r) \cap X_{\{n\}}\right| \geq k\right\}$, which is the distance from x to its k-th nearest neighbor. Here we take advantage of k-NN estimator [48] as

$$f_k(x) := \frac{k}{n \cdot v_d \cdot r_k(x)^d} \tag{10}$$

n stands for image pixels. v_d is the volume of a unit ball in feature space. k is the number of neighbors that a point x has. Suppose that f has compact support χ. Therefore, cluster-tree estimation is done by k-NN graph to determine the level-sets of f. We also define $G(\lambda)$, which denotes the mutual k-NN graph under the condition of

$$f_k(x) \geq \lambda \tag{11}$$

$$\left|\left|\mathrm{x}-\mathrm{x}'\right|\right| \leq a \cdot \min\left\{r_k(x), r_k\left(x'\right)\right\} \ (a=1) \tag{12}$$

It approximates the λ-level set of f_k. λ is referred as a corresponding empirical density to determine clusters instead of radius r, which has only difference in form. When λ approaches zero, not only the connected components set (CCs) but also connectivity between nodes appear or merge at every level λ. CCs also forms a hierarchical nesting structure by performing the top-down sweep of k-NN estimation and each one corresponds to its local maxima of f, thus a disjoint modal-set of χ comes in to being without overlap.

According to what we discussed above, we pinpoint the cluster core M as

$$x \in \chi : f(x) \geq \alpha \cdot max_{x' \in M} f\left(x'\right) \tag{13}$$

The range of α is from 0 to 1. Note that when $\alpha \to 1$, M stands for the local maxima of f while $\alpha \to 0$, it becomes the entire χ. It is just what we make a slight difference from Jiang's method [47]. What he proposed is using a fixed additive fluctuation β to check level $(\lambda - 9\beta_k\lambda)$ from M nearby and picking up the points as estimation. It is conducive to determine the right number of CCs since there may be various modal-sets at some level λ. However, it requires knowledge of the scale of the density function, which is too tough to obtain in practice. Hence, we choose a multiplicative fluctuation parameter α to take the place of Jiang's fixed parameter, which adapts to clusters at different density levels more appropriate. So in the case that the CCs appears in $G(\lambda)$, we can take the corresponding CC in G($\alpha \cdot \lambda$) to determine how much the density is allowed to vary and estimate the cluster-core accurately.

On successfully obtaining the cluster cores as an initialization, quick shift is then run based on hill-climbing procedure on each remaining sample and each sample is moved to its nearest neighbor that has a higher k-NN density until it reaches a cluster-core. Samples that end up in the same cluster-core, or even the ones outside the cluster-core are guaranteed to be assigned to the same cluster. Aiming at close-shots images, we refer to the method proposed by Winn J [49] to extract the feature. The filter bank consists of three Gaussian kernels (with $\sigma = 1, 2, 4$ applied to each CIE L, a, b channel), four Laplacians (with $\sigma = 1, 2, 4, 8$ applied to L channel) and four derivative Gaussians (divided into x- and y-aligned sets with $\sigma = 2, 4$). Totally, a 17-dimensional feature vector signifies each pixel of the image, making the words more diversiform. Meanwhile, it enhances the feature description ability along with a robust convergence of the unsupervised word topic learning. By sampling from the document distribution of each pixel according to the modified quick shift, a reasonable coding for spatial structure can be achieved. Likewise, given this word document assignment, the probability distribution for each pixel can be finally obtained by LDA. We simplify the prior steps below:

- Initialize the cluster-core $M = \varnothing$, $k = 100, \alpha = 0.1$. Input the image data and operate k-NN density estimation. Sort x_i as the rule $f_k(x_i) \geq f_k(x_{i+1})$;
- Define $\lambda := f_k(x_i)$, find disjoint connected components set (CCs) of G($\alpha \cdot \lambda$) containing x_i and add them into M;
- Initialize directed graph G with vertices of image data $\{x_1, \ldots x_n\}$ and no edge. Conduct quick shift process and judge whether x_i is in any cluster-core. If not, add an edge to G from x_i to its nearest sample x under the condition of $f_k(x_i) \geq f_k(x_{i+1})$.
- Find $\hat{C_M} \in M$ where x has the directed path in G starting at x ends in M and allocate $\hat{C_M}$ as the documents of LDA. Obtain the 17-dimensional feature vectors with the document collection $\hat{C_M}$ to execute the LDA process. Compute the topic word parameter matrix by Gibbs sampling and classify the image into K parts according to the probability.

In this way, we design the document allocation for close shots properly. On the one hand, it is internally effective to conduct image analysis by mode seeking aggregation and nonparametric density estimation. It replaces the rigid structure of pixel grid with meaningful atomic regions and preserves the valid boundary information of objects in the image. These regions of various shapes can also allow documents to be more flexible in number of words based on practical spatial construction.

In this way, it can contribute to the inclusiveness and accuracy of probability distribution form among topics in each document. On the other hand, this modified quick shift algorithm has a good global generalization ability through the kernel density estimation with smooth filtering effect. It is possible to model the rich variety of locally high-density structure in complex image data, resulting in greatly improved accuracy of LDA.

According to the discussion above, though it is of little possibility for LDA to classify every pixel precisely, SA-LDA based on statistical model can acquire a relatively reliable labeling result efficiently. Through preliminary analysis of the image type, we determine an adaptive method for two different kinds of pictures. As a result, it can automatically select the appropriate feature and distribute documents according to the corresponding image characteristics.

2.4. Self-Adaptive Iterative Latent Dirichlet Allocation (SAI-LDA)

As a matter of fact, our method is finally applied to the practical greenhouse environment, which means that we still need to analyze for the specific situation. Sometimes even human will make mistakes in distinguishing ropes, reddish-brown soil and rusty pipes from fruits or leaves since they seem to be exactly the same due to the low quality of images. Besides, images with tomatoes at different cultivation periods are all included in our dataset, which means the difference of their color, shapes, and density should be taken into consideration. There are also new leaves, curly old leaves, and diseased leaves with yellow spots existing in images. How to balance the inclusiveness of the characteristics of the same class of objects poses a great challenge for us. Therefore, we need to make further efforts to improve the SA-LDA for better adaption for more images.

In Section 2.3.2, we choose the wavelet feature extraction for distant shots, which is also a hierarchical method to compensate for the shortcoming of single resolution. We found that better results can be obtained with the multiresolution decomposition of the original image [50–52]. Among these hierarchical algorithms, image pyramid is the most common structure, the theory of which has been put forward and developed for several decades. The image is divided into several resolutions by down-sampling the original image level by level with the factor 2i (i = 1, 2, 3 . . .), forming a successive collection like pyramid structure. The process can be terminated at any level depending on precision requirements and fitness of different methods. The pyramid structure can be seen in Figure 13.

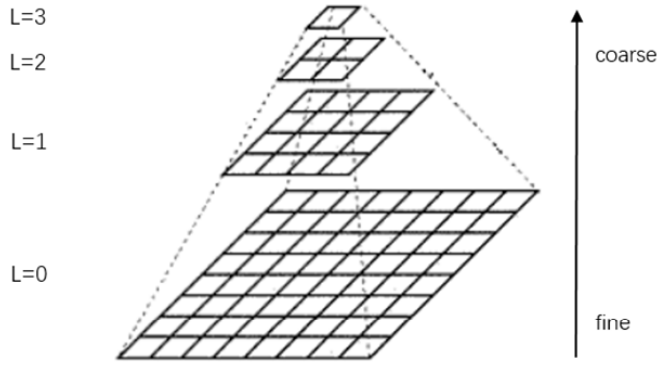

Figure 13. The structure of the image pyramid. L stands for the image layer, from the original (L = 0) to coarser ones by continuously down-sampling.

Referring to the ideas above, we can similarly perform multi-resolution operation at the level of document allocation, serving as a supplement to combat the drawback of SA-LDA. To this regard, we can see that the same size of patches in different level can include different range of feature. For example, a 50 × 50 image window can acquire a whole tomato at the highest resolution while only a small part of fruit at the lowest one. Furthermore, the incompatibility of tomato characteristics in different morphology and illumination influence can be alleviated due to the fact that features of the reflective or shadowed part will be distinctly blurred to a certain extent by decreasing the resolution

layer by layer. Therefore, we generate this image pyramid to explore richer image information and sample more words through several levels of documents. Based on this hierarchical idea, our optimized SA-LDA algorithm can well combine the bag-of-words model with multiresolution theory, which shows its superiority on executing more precise segmentation by its iterative process. We abbreviate this iterative method as SAI-LDA, of which the procedure is as follows.

Our original image has the size of 200 × 300 pixels (or 300 × 200 pixels) and we divide the image into three levels on the basis of experimental experience. Down-sampled twice in orthometric directions, two layers of images are acquired. Level 0 corresponds to the original image. Level 1 is one-half while level 2 is exactly one-quarter of the preliminary image. According to our discussion, we carry out SA-LDA proposed in the previous chapter on image at level 2 as the first step of LDA document iteration to make a judgment of the image type. Different types of images result in their corresponding feature extraction and document allocation. Therefore, we have the access of the first segmentation result. Then we enlarge the result by a factor of two to level 1 and regard it as the initial document allocation of the next LDA process, which is cycled to obtain the second LDA segmentation output. Afterwards the second output is enlarged again to be the last document allocation and finally provide classification result. In this way, the hierarchical method can compensate for the effect of illumination and shadow. Besides, some differences between the same class can be weakened, hence it improves the robustness and accuracy of LDA from the aspect of feature extraction and document allocation, facilitating to the subsequent description of spatial information. That is why finer results can be obtained than single resolution through multiple cyclic iterations. Our algorithm flowchart is shown in Figure 14.

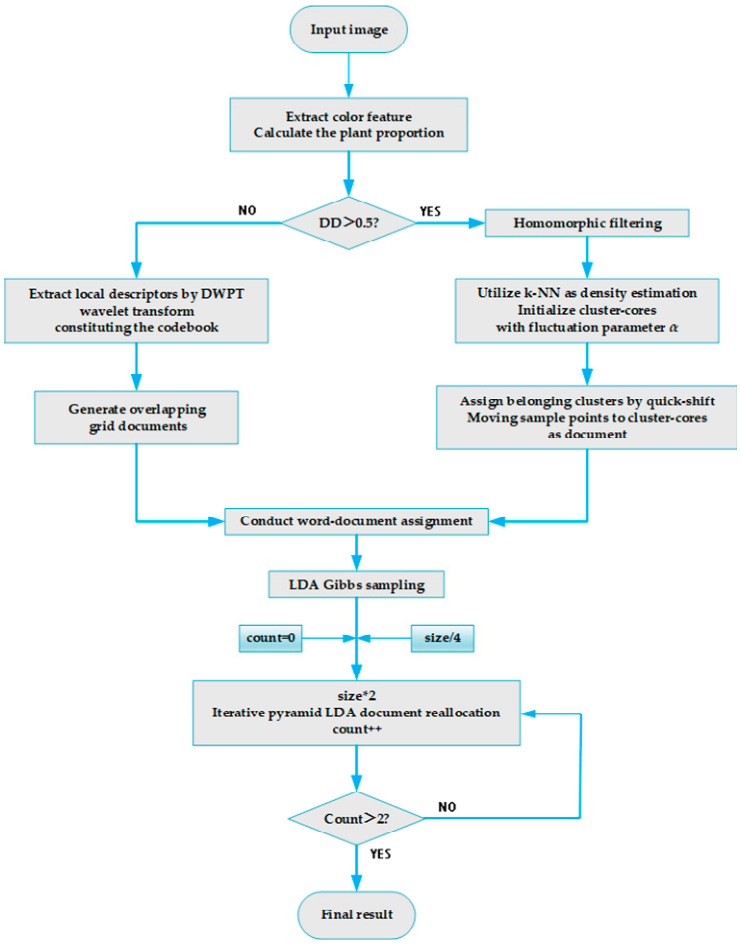

Figure 14. The flowchart of self-adaptive iterative latent Dirichlet allocation algorithm (SAI-LDA). DD is the distance determination; DWPT is discreet wavelet packet transform.

3. Results

3.1. Massive Data Acquisition Based on The Fixed Camera

Due to our ultimate goal for the establishment of database for tomato plant organs, a stable data acquisition source is needed as a premise. In other words, tomatoes of all growth stages, from seedling to fruit setting should be included in our database. Photographs taken by human are not only difficult to meet the research needs in quantity, but also cannot fully display the plant morphology of different periods or angles. The introduction of automation and digital imaging allowed the rapid collection of real-time images of plants in a non-destructive fashion. Therefore, we built a camera image acquisition system in the glass greenhouse of Chongming National Facilities Agricultural Engineering Technology Research Center in Shanghai and transmit tomato plant images to the laboratory in real time. Such systems are simpler to deploy, more affordable, and have higher throughput since they can image many plants at a time. The greenhouse tomato plants are arranged in rows. About 25 tomato plants are cultivated in each row with an interval of about 80 cm between rows, as shown in Figure 15. Considering many factors such as physical equipment interferences, the greenhouse layout, network environment and so on, we installed a 2-megapixel spherical infrared night vision network surveillance camera on the wall bracket of the glass greenhouse. It can output real-time images of 30 FPS at the highest resolution and automatically save images to the memory card. The installation position is shown in the red circle in Figure 15a.

Figure 15. The greenhouse environment. (**a**) The installation site of our fixed camera; (**b**–**d**) different growth stages and shooting angle of tomato plants.

The camera has the capacity of self-head-rotating in 360 degrees continuously in horizontal direction and 90 degrees motion range vertically. Therefore, we can design the camera scanning trajectory and set fixed location to achieve multi-angle shooting of plants depending on our experimental requirements. Aiming at shooting tomato fruits, we initially selected 200 locating points with different focal lengths. Firstly, we set the initial focal length to 50 mm and made our camera moving from south to north according to the cruise trajectory, as shown in Figure 16. We are aware of the fact that the leaf growth in middle and higher parts of the whole plant is almost the same. Furthermore, the vertical rotation angle of the camera is unfortunately limited due to the mechanical

construction of our camera. To this regard we set the cruise trajectory only to make sure that complete tomato fruits are available in images on its height, except the first and the last row. As for the shooting frequency, the head rotation of our camera shifts every 15 s, taking two or three pictures at each location to avoid camera shake. After accomplishing the shooting task of a complete cruise route, a new round of trajectory is carried out by setting a shorter focal length of 10 mm. Note that the target will be blurred due to extremely short focal length, we set the final minimum focal length to 10 mm. In this way, we are able to obtain distant or close shots of tomato plants. Our shooting task in one day is conducted on three certain periods: 8:00 a.m. to 9:00 a.m., 12:00 a.m. to 1:00 p.m., and 4:00 p.m. to 5:00 p.m. The strategy of taking pictures in separated 3 h instead of constantly shooting for a long time is conducive to obtain more kinds of images under different lighting conditions, which enhances image diversity in our database. In theory, about 1440 to 2160 images can be taken in a day and we name these images according to each orientation and focal length to set up the initial data source. In fact, the morphology of tomato plant will not change macroscopically in a short term, so we take pictures every five days to save resources as well as to guarantee the image quantity. When it comes to our further experiment, we condense them to 200×300 (or 300×200) for convenience of algorithm execution.

Figure 16. The cruise trajectory (red line with arrows) of our shooting strategy: moving up and down with varied focal length from the first row to the last.

3.2. Comparison Settings and Results

Here we provide some experiments to verify the superiority of our algorithm, comparing with several state-of-the-art approaches utilizing quantum-behaved particle swarm optimization (QPSO) for muti-threshold searching [20], fuzzy c-means clustering (FCM) [16], pulse coupled neural network-based segmentation (PCNN) [14], Co-segmentation [53] and the traditional LDA, seen from Figures 17–20. We use red, green, and blue color to represent the labeled area of fruits, leaves, and the background, respectively. In addition, high quality ground-truth images are also demonstrated as segmentation criterion of all the results. The analysis of theses comparison experiments can be seen in Section 4.1.

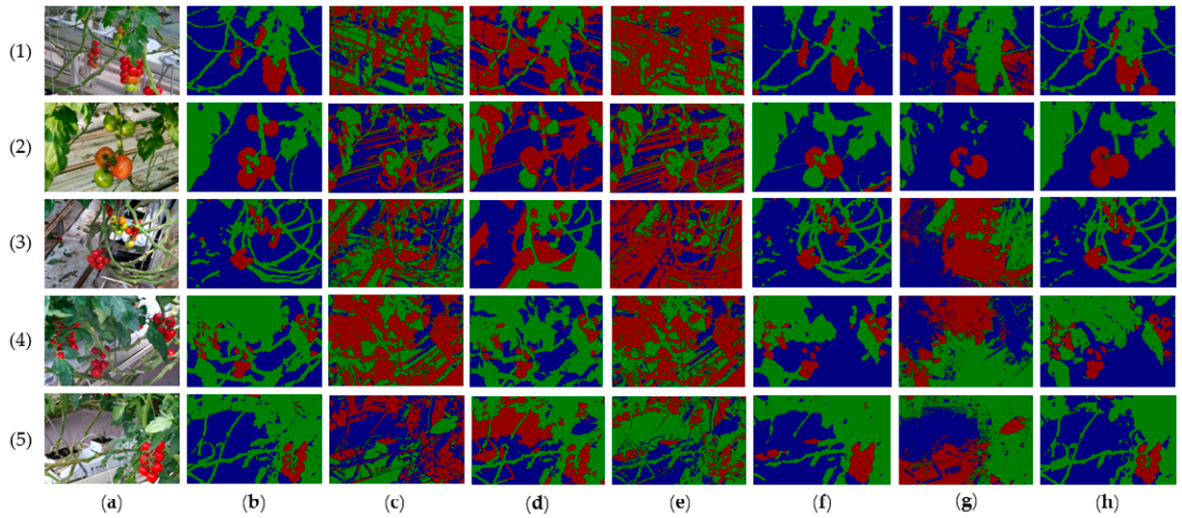

Figure 17. The result of different algorithms with distant shots. Column (**a**) are original images. Column (**b**) are high quality ground-truth. Columns (**c**–**h**) are the results of several segmentation methods: quantum-behaved particle swarm optimization (QPSO), fuzzy c-means clustering (FCM), pulse coupled neural network-based segmentation (PCNN), co-segmentation, the traditional latent Dirichlet allocation (LDA), and self-adaptive latent Dirichlet allocation (SA-LDA), respectively.

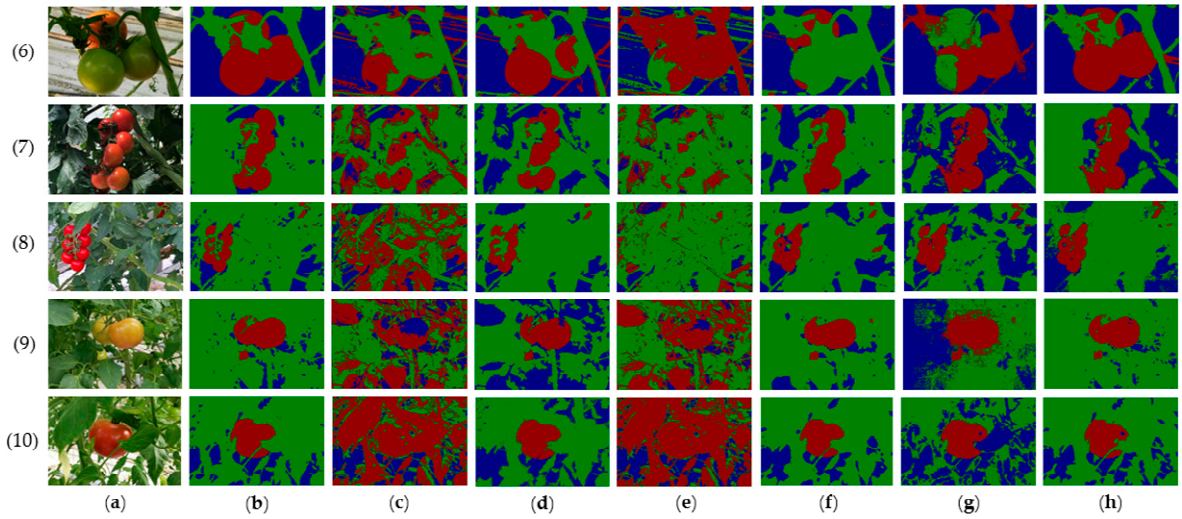

Figure 18. The result of different algorithms with close shots. Column (**a**) are original images. Column (**b**) are high quality ground-truth. Columns (**c**–**h**) are the results of several segmentation methods: quantum-behaved particle swarm optimization (QPSO), fuzzy c-means clustering (FCM), pulse coupled neural network-based segmentation (PCNN), co-segmentation, the traditional latent Dirichlet allocation (LDA), and self-adaptive latent Dirichlet allocation (SA-LDA), respectively.

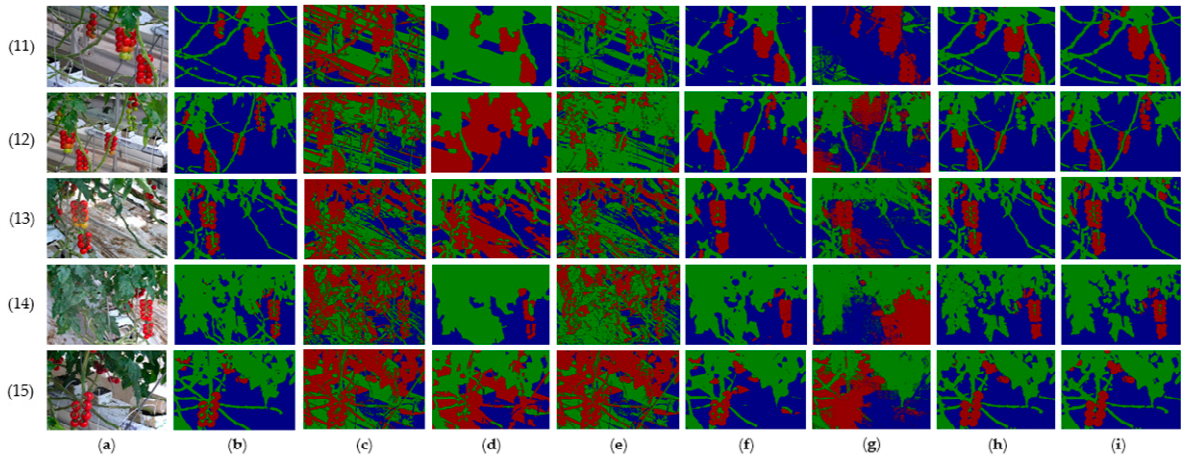

Figure 19. The result of different algorithms with distant shots. Column (**a**) are original images. Column (**b**) are high quality ground-truth. Columns (**c**–**i**) are the results of several segmentation methods: quantum-behaved particle swarm optimization (QPSO), fuzzy c-means clustering (FCM), pulse coupled neural network based segmentation (PCNN), co-segmentation, the traditional latent Dirichlet allocation (LDA), self-adaptive latent Dirichlet allocation (SA-LDA), and self-adaptive iterative latent Dirichlet allocation (SAI-LDA), respectively.

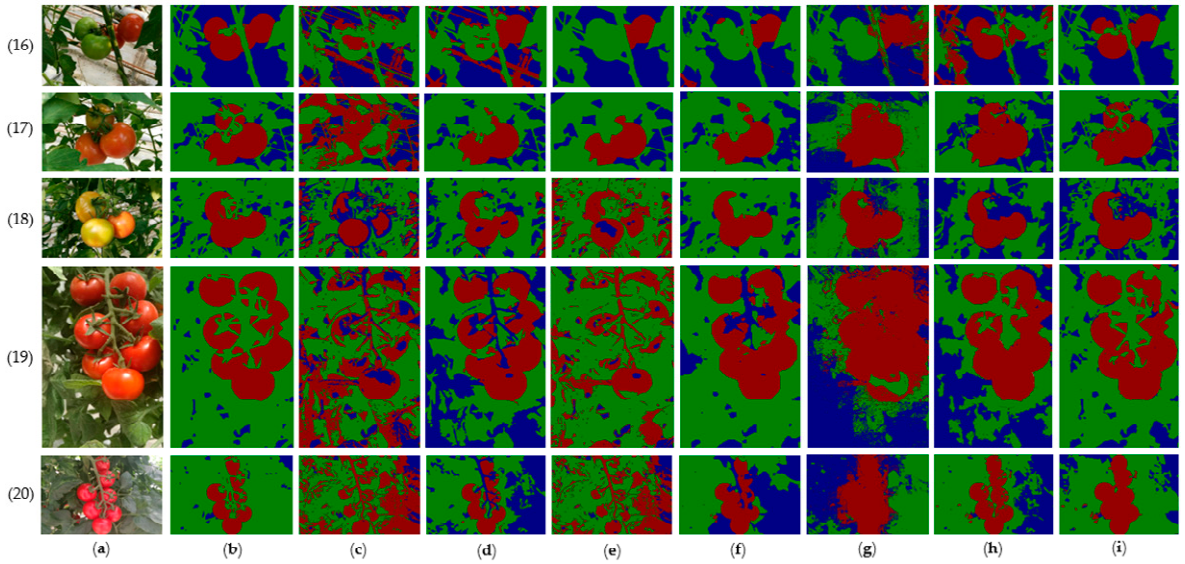

Figure 20. The result of different algorithms with close shots. Column (**a**) are original images. Column (**b**) are high quality ground-truth. Columns (**c**–**i**) are the results of several segmentation methods: quantum-behaved particle swarm optimization (QPSO), fuzzy c-means clustering (FCM), pulse coupled neural network based segmentation (PCNN), co-segmentation, the traditional latent Dirichlet allocation (LDA), self-adaptive latent Dirichlet allocation (SA-LDA), and self-adaptive iterative latent Dirichlet allocation (SAI-LDA), respectively.

3.3. Evaluation Criterion

Our evaluation technique is based on the comparison of the segmentation results and ground-truth by human labeling judging from two aspects. One is the pixel accuracy of the overall image (IA), taking fruits, leaves, and the background together into consideration. The other is the accuracy of the plant organs respectively—that is, fruit accuracy (FA) and leaf accuracy (LA)—due to our goal of distinguishing them from the image.

Here we define

$$\mathrm{IA} = \frac{P_c}{m \times n} \tag{14}$$

$$\mathrm{FA} = \frac{P_{cf}}{P_{cf} + P_{ff}} \times 100\% \tag{15}$$

$$\mathrm{LA} = \frac{P_{cl}}{P_{cl} + P_{fl}} \times 100\% \tag{16}$$

P_c is the number of pixels that have the correct label compared with the ground-truth; m is the width and n is the height of the image so that $m \times n$ signifies the overall number of pixels. P_{cf} stands for the number of the correct pixels for fruits while P_{cl} stands for those for leaves. P_{ff} is the number of labeled pixels for fruits in our experimental result but are actually wrong in the ground-truth. Similarly, P_{fl} is the number of false pixels for leaves appearing in our image.

We also define precision (P), recall (R), F_1 (F-measure) to form our final pixel-based evaluation criterion. Those are given by

$$\mathrm{P} = \frac{1}{K} \sum_{i=1}^{K} \frac{P_i^c}{P_i^c + P_i^f} \tag{17}$$

$$\mathrm{R} = \frac{1}{K} \sum_{i=1}^{K} \frac{P_i^c}{P_i^c + P_i^m} \tag{18}$$

$$F_1 = \frac{(a^2+1)PR}{a^2(P+R)} = \frac{2PR}{P+R} \tag{19}$$

P_i^c is the number of correct detected pixels corresponding with the ground-truth, P_i^f is the number of false detected pixels, P_i^m is the number of omissive pixels which exist in ground-truth but are missed in the experiment result. K is the number of classes representing for fruits, leaves and the background, and hence K = 3. The precision and recall are both effective criteria to evaluate segmentation quality, that is, the higher the parameter values are, the better the segmentation result is. Moreover, the combination of the two measures can both indicate over- and under-segmentation, and thus present an integrated result. However, these two parameters in some cases will act in constraint with each other so that we turn to F-measure to take into account the joint measure of precision and recall. It captures the trade-off and forms a harmonic mean between the two. Here we choose parameter a = 1 according to experience. F_1 is in high value if both precision and recall are high; however, if one of them has low value, the value of F_1 will go down.

Tables 1 and 2 show the comparison of accuracy among different methods. The highest accuracy rates are in bold.

Table 1. Image segmenting accuracy (IA) of the comparative experiments. The methods below are quantum-behaved particle swarm optimization (QPSO), fuzzy c-means clustering (FCM), pulse coupled neural network-based segmentation (PCNN), co-segmentation, the traditional latent Dirichlet allocation (LDA), and self-adaptive latent Dirichlet allocation (SA-LDA).

	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
QPSO [20]	0.2307	0.5502	0.1900	0.1125	0.0814	0.5238	0.6404	0.6159	0.6302	0.2901
FCM [16]	0.4671	0.7105	0.5228	**0.8249**	0.7002	0.7811	0.6803	0.8319	0.6994	0.8647
PCNN [14]	0.1544	0.3515	0.1502	0.2766	0.3079	0.7662	0.5400	0.7742	0.6006	0.2340
Coseg [53]	0.8654	0.8167	0.8833	0.5802	0.7034	0.3943	**0.7290**	**0.8612**	0.8604	0.8849
LDA	0.5793	0.6704	0.3743	0.3938	0.6003	0.6819	0.4451	0.7028	0.5709	0.6563
SA-LDA	**0.9533**	**0.9236**	**0.9391**	0.7669	**0.8766**	**0.9526**	0.6367	0.8001	**0.9122**	**0.9113**

Table 2. Image segmenting accuracy (IA) of the comparative experiments. The methods below are quantum-behaved particle swarm optimization (QPSO), fuzzy c-means clustering (FCM), pulse coupled neural network based segmentation (PCNN), Co-segmentation, the traditional latent Dirichlet allocation (LDA), self-adaptive latent Dirichlet allocation (SA-LDA), and self-adaptive iterative latent Dirichlet allocation (SAI-LDA).

	11	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**
QPSO [20]	0.1860	0.2044	0.1557	0.1239	0.1714	0.6411	0.4919	0.8203	0.8316	0.7330
FCM [16]	0.3372	0.3816	0.4507	0.8403	0.6218	0.6629	0.8417	0.8656	0.7227	0.7048
PCNN [14]	0.2105	0.1789	0.1919	0.1046	0.1599	0.7405	0.8008	0.7744	0.6848	0.7016
Coseg [53]	0.7954	0.8883	0.8612	0.8682	0.8309	0.7782	0.8311	**0.8820**	0.8109	0.5528
LDA	0.6498	0.1099	0.7526	0.6503	0.6455	0.7653	0.4856	0.6514	0.4813	0.4963
SA-LDA	0.8855	0.9018	0.8993	0.8336	0.8875	0.8242	0.8725	0.7484	0.7402	0.7349
SAI-LDA	**0.9485**	**0.9589**	**0.9411**	**0.8737**	**0.9280**	**0.9474**	**0.9384**	0.8321	**0.8936**	**0.8421**

The scatter diagram of fruit or leaf accuracy are shown in Figures 21 and 22. Figure 23 expresses the F_1 value of the comparative experiments.

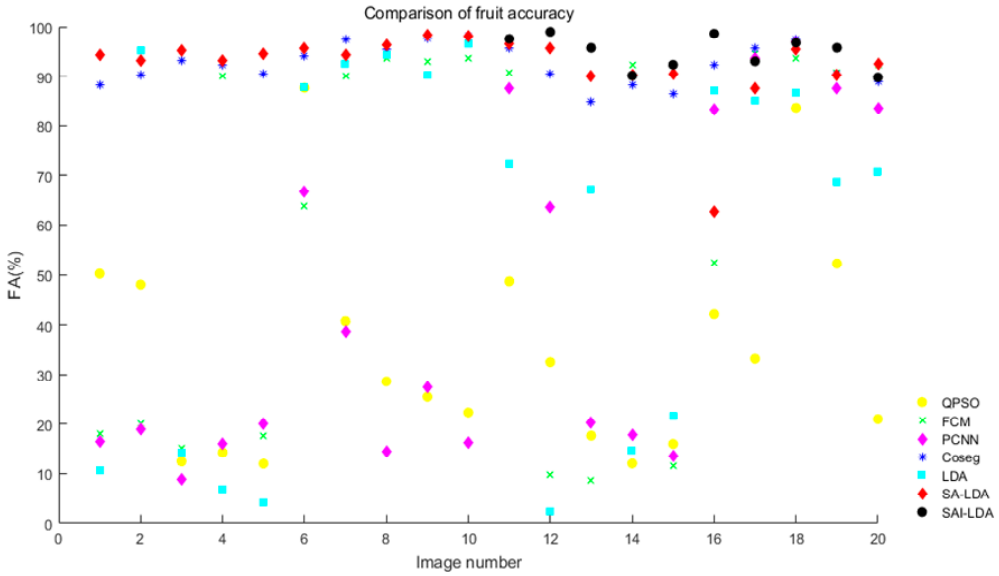

Figure 21. The result of fruit accuracy with different algorithms.

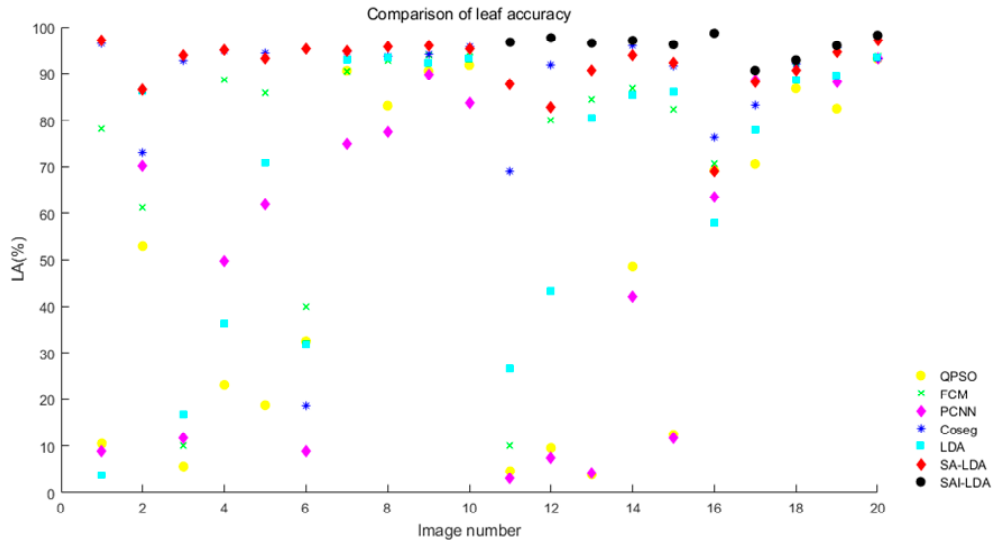

Figure 22. The result of leaf accuracy with different algorithms.

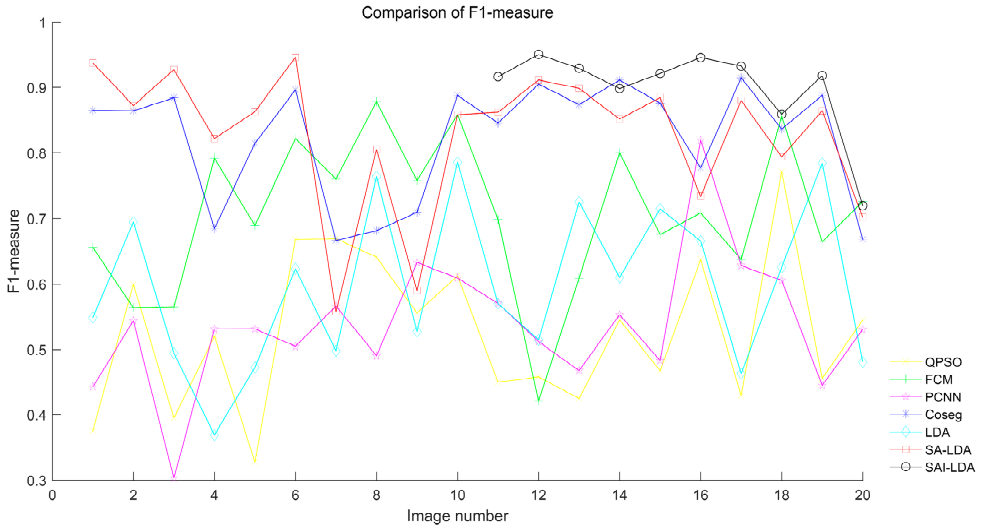

Figure 23. Comparison of F_1-measure value.

4. Discussion

4.1. Analysis of Comparative Experiments

We selected three unsupervised segmentation algorithms [14,16,20] in the field of plant segmentation for comparative experiments. Through the results, we can see that when the same segmentation algorithm is applied to diverse backgrounds, their weaknesses will emerge at different level. In [20], quantum-behaved particle swarm optimization (QPSO) is actually a heuristic method for multi-threshold random search optimization. The pixel value is regarded as a particle and its position is constantly updated, combined with the spatial information of images. In this way it can maximize the fitness function according to the difference between classes and gradually approach the global optimal threshold. The multi-state characteristic of a quantum system enables itself to search any position in the whole image space with probability in a certain range, which has better flexibility, convergence, and stability. For images in our database, we set the number of particles to 150 and the maximum number of iterations to 200, adopting the minimum error function of the ratio of variance between classes as the fitness function. We found that the results are quite unsatisfactory, especially in distant shots. QPSO shows light-sensitive results and often confuses fruits and leaves, identifying the whole plant as one class, as shown in image 2, 3, 5, 13, and 15. It ascribes to the fact that QPSO directly entrusts boundary information to particles when particles fly out of the boundary of the search area, without considering that the boundary may also produce local optimal solution. In practice, the global optimal solution is unlikely to be located at the boundary of the search area while the boundary between different classes in our data is too obscure for it to obtain better optimization. To this extent, more and more particles gather at the boundary and the algorithm falls into local optimal situation.

Artificial pulse coupled neural network PCNN is used in [14] for the recognition and segmentation of fruits. It matches the pixels with neurons and connects each other to form a single-layer, two-dimensional local feedback network. It does not need training and utilizes different intensity of pixels to input external stimulus, conducting igniting process at all times. Neurons (pixels) with similar characteristics can generate pulses at the same time so that they can compensate for the spatial discontinuity or small changes in the amplitude of the input data and preserve the regional information completely. In the experiment, we selected (1.3R-G-B) as the fruit stimulus input and (R-1.5B) as the leaf stimulus input according to our previous conclusion in Section 2.3.1, and make class judgement of each pixel by ignition respectively. When a pixel is judged as either fruit or leaf simultaneously in two processes, we endow it with the same attribute depending on its eight neighborhood pixel categories. We set the attenuation coefficient α = 0.1, the connection coefficient β = 0.2, and the weight matrix W = [0.707, 1, 0.707; 1, 0, 1; 0.707, 1, 0.707]. Due to the inherent biological mechanism of PCNN, when

minimum cross-entropy is used as the adaptive iterative control criterion, the high complexity of the image and the uneven illumination lead to multi-peak distribution on the gray level. That is why poor detail performance occurs with over-smoothing and cross-blurring. It can be proved that the algorithm can hardly achieve reasonable multi-target segmentation for our images.

Tomato fruit recognition based on the principle of fuzzy c-means proposed in [16] is a "soft assignment" partition clustering method and pixels can belong to multiple classes with different membership degrees. It retains more information from the original image with the lowest similarity between classes and the greatest one within classes by the objective function. Although the concept of neighborhood density is introduced to initialize the clustering center in the literature and it does show higher segmentation accuracy than the above two, however, FCM is effective only for the recognition of mature tomatoes in close shots similar to [14]. As shown in image 4, 7, 8, 10, 17, 19, and 20, all tomatoes have very high visual saliency in the images. In our database, the existence of green tomatoes and leaves under a busy background in different growth cycles is common, which brings great interference to the initial clustering center in the color space. Meanwhile, the different proportion of tomatoes in the image also causes fluctuations in the performance of the algorithm, which results in the decrease of algorithm robustness.

We also adopt the traditional LDA algorithm for contrast experiment to verify our breakthrough in the internal mechanism of LDA. We use the feature extraction method introduced by Winn J [49] and choose the rectangular non-overlapping grids for document allocation. The size of the grids are 50×50 pixels. Note that the lower resolution the image has, the more noise it will produce. The document allocation strategy of the traditional LDA is rough and simple, and words can simply describe the details of the image rather than the global information or the relationship between objects. Thus, this mapping of the documents does not have the capacity to reflect the spatial structure of an image to some extent. For images 3, 4, and 5, there are obvious reflection problems while for images 19 and 20, different parts of leaves show various shades due to light and occlusion. As the fruits and leaves occupy a larger proportion of the whole close-shot images than the distant ones, each non-overlapped 50×50 patch contains a smaller feature difference in the specific organ. Words that are at a relatively distant distance cannot be included in a document so it will be less accurate under LDA algorithm with this kind of word document construction. Meanwhile, the poor quality of images also makes it difficult for documents to tolerate feature differences on the same object, resulting in the lack of global generalization of the algorithm.

As for the last algorithm co-segmentation, it is a weak supervised algorithm that needs to build up a training set to distinguish the common object in the image group. Instead of just focusing on regions with uniform local features, it takes into full account of semantic information of the image data to execute discriminant clustering, classification learning, or modeling optimization. In this way it can generally achieve better segmentation. During the experiment, we use 30 images as an input to build a training set and produce corresponding results. However, the ability of discovering local aspects of color, texture, and shape features is impeded by the sub-optimal images. The unary and pairwise energy potentials across all the images, which can transfer the co-occurrence patterns from the object proposals to each image are also difficult to define. Furthermore, the more pictures in the training group, the longer learning time it will take, which runs counter to our goal in quick labeling. Although it seems to be more suitable in some types of images, this weak supervision method does sacrifice time complexity with only a little accuracy improvement and hence does not serve as an appropriate algorithm in reality.

4.2. Analyzation of Proposed Algorithm

When it comes to our SA-LDA algorithm, firstly it divides images into two categories, one is close shots, the other is distant shots. According to the adaptive document allocation and hierarchical feature extraction, we can effectively overcome the shortcomings of the methods mentioned above. The highest F_1 value, FA and LA can reach 0.9533 (image 6), 0.9774 (image 9) and 0.9411 (image

1), respectively. Compared with the traditional LDA, our dynamic document design is no longer restricted by the rigid structure. Instead, it considers more from the characteristics of the image data itself, leading to smoothness of the shape and edge. In addition, the use of merging pattern points for documents can effectively reduce the noise brought by the BoW model and over-segmentation by single point-mode, confirming the excellent global generalization ability of SA-LDA. Through our experiments data, SA-LDA hits an excellent grade in comparison with other algorithms in greenhouse environment, especially for close shots with varied illumination intensity or distribution as well as pictures consist of some vine-like stents (image 1), rusty pipes (image 2, 3, 6, 16), the ground or soil stains (image 13). These disturbance factors can also be well distinguished from tomato plants and then be divided into background class, providing a more exact, semantically segmented result.

Nevertheless, we also find some drawbacks of SA-LDA indeed. In image 1, 2, 11, and 12, some immature green tomatoes are still misclassified as leaves. This is due to the fact that small difference between green fruits and leaves cannot be completely distinguished only by the wavelet iteration from feature space. Furthermore, in image 14, the algorithm is easily disturbed by lesion, which is the large area of bright yellow spots on the leaves, thus it is mistakenly classified into the background. The uneven illumination, the cluttered leaves, the shadows caused by overlapping fruits in image 7, 9, and 10, have hindered the improvement of accuracy.

Therefore, on the basis of the adaptive LDA algorithm, we consider a pyramid structure from low resolution to high resolution to modify the document allocation. Seen from the last column in image 14, most of the leaf lesions are classified into the area of correct leaves instead of the background. Through the adaptive LDA at the lowest level, we get a better initial segmentation, which can weaken the characteristics of the lesions to a certain extent, so that the difference of illumination or lesions on the organ itself has less holdback on feature extraction. The influence of feature differences on the same class of objects is reduced by cumulative steps. At the same time, green fruits are well recognized, as shown in image 11, 12, 13, 16. The FA value also increases about 5.02% on average. Although at the lowest resolution level, the differences between green fruits and leaves are also weakened, the experimental results verify that the weakening effect is relatively slight so that the uniformity of the illumination brought by several iteration prevails. It is apparent that during subsequent iterative process, the algorithm not only continues to make judgments and balance weights among green items, but also makes up for some defects, including the lack of description of local details and sensitivity to feature differences generated on each resolution. Therefore, the higher accuracy, tolerance, and robustness make SAI-LDA surpass SA-LDA in many cases.

In a word, the segmentation results of the upper layer, which is regarded as the initial document allocation, show great impact on labeling results of the next layer. Meanwhile, the final results can be affected through three rounds of iteration layer by layer. We can verify that SAI-LDA has two improvements as breakthrough towards the former SA-LDA: one is to solve the inclusivity of old, infected leaves and normal leaves. It adapts well to uneven lighting and reflection problem, which are common in the greenhouse. The second is to precisely tell the similar green fruits and leaves apart under suboptimal circumstances through several iterations, which strikes a balance between co-occurrence information and feature differences within the spatial structure, leading to accurate organ identification.

When it comes to the time dimension, the data in Table 3 demonstrates that the average running time of SA-LDA is 35.9% less than that of the traditional LDA. Likewise, the running time of SAI-LDA is 30.3% less than that of SA-LDA. It is apparent that our method can not only improve the calculation speed but also increase the accuracy by modifying the document allocation and conducting hierarchical LDA iteration to tackle the drawbacks of the time-consuming LDA method. In general, this unsupervised algorithm can label the tomato plant organs fast and effectively so that it is conducive to subsequent image processes.

Table 3. Average time consumption between seven algorithms. They are quantum-behaved particle swarm optimization (QPSO), fuzzy c-means clustering (FCM), pulse coupled neural network based segmentation (PCNN), a weakly supervised algorithm co-segmentation (Coseg), the traditional latent Dirichlet allocation (LDA), self-adaptive latent Dirichlet allocation (SA-LDA), and self-adaptive iterative latent Dirichlet allocation (SAI-LDA) (unit: second).

Methods	QPSO	FCM	PCNN	Coseg	LDA	SA-LDA	SAI-LDA
Time Consumption (s)	1.72	7.66	52.54	218.65	131.64	84.39	58.79

5. Conclusions

In this paper, we propose a modified statistical model of LDA, namely SAI-LDA to conduct tomato plant segmentation based on the greenhouse camera images, which is carried out in an entirely unsupervised way. Through our experiments in different cases, some conclusions are drawn as follows:

1. Firstly, we analyze the unsatisfactory image quality from camera itself and the complexity of the greenhouse environment, explaining the limitations of the traditional LDA algorithm in such situation. Combining the significance of feature extraction and spatial structure of the original data, we first bring out a self-adaptive algorithm SA-LDA, where different pixel characteristics and spatial coding criteria rely on distant shots or close shots. The spatial structure of the image can be judged according to score of DD formula, that is a calculation of the proportion of plants based on color characteristics of fruits, leaves, and the background, respectively. For distant shots, the improved DWPT wavelet feature extraction is adopted, which makes use of the standard deviation and energy of LL, LH, and HL wavelet sub-bands together with CIE Lab color feature. Meanwhile, overlapping rectangular document allocation is conducted to enhance the traditional LDA algorithm. Given the fact that close shots are easily affected by illumination variation, the homomorphic filtering serves as a preliminary. Afterwards, we innovatively use modified quick shift to obtain multiple density-based mode points, which is non-parametric estimated, and then form modal-sets as clusters to flexibly execute document allocation.

2. Aiming at the problems still existing in SA-LDA, including insufficient inclusiveness of features of the same object and the lack of discrimination of objects with strong similarity, a pyramid iteration method namely SAI-LDA is further proposed. Divided into three layers of pyramids, we adopt segmentation result of SA-LDA as the initial document starting from the lowest resolution. The outputs of each layer are enlarged and utilized as the initial document allocation of the next higher resolution. The final results are obtained twice by iteration, which is finer in the spatial structure of documents than SA-LDA with optimal labeling results and less consuming time. The experimental results show that our algorithm overcomes the objective problem of low quality and random structure of complex greenhouse plant pictures from a fixed camera and completes the identification of plant organs automatically. It has the capacity of serving as a visual phenotype information complemented with other sensors in greenhouse as well. Thus, it contributes to laying a foundation for mass acquisition of training samples for future work, which has a great practical significance in the field of agriculture.

Author Contributions: Q.C. set up the image acquisition system in the greenhouse, made investigation, conceived and performed the experiments, conducted formal data analysis, prepared figures and/or tables, wrote the manuscript of the paper, approved the final draft. L.X. established the guidance for the research idea, writing of the manuscript, authored or reviewed drafts of the paper, approved the final draft.

Acknowledgments: The authors are greatly thankful to support by National Natural Science Foundation of China (grant No. 61573258), National High-Tech R&D Program of China (grant No. 2013AA102305) and US National Science Foundation's BEACON Center for the Study of Evolution in Action (DBI-0939454).

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Sodhi, P.; Vijayarangan, S.; Wettergreen, D. In-field segmentation and identification of plant structures using 3D imaging. In Proceedings of the 017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Vancouver, BC, Canada, 24–28 September 2017; pp. 5180–5187. [CrossRef]
2. Zhang, C.; Si, Y.; Lamkey, J.; Boydston, R.A.; Garland-Campbell, K.A.; Sankaran, S. High-Throughput Phenotyping of Seed/Seedling Evaluation Using Digital Image Analysis. *Agronomy* **2018**, *8*, 63. [CrossRef]
3. Tripodi, P.; Massa, D.; Venezia, A.; Cardi, T. Sensing technologies for precision phenotyping in vegetable crops: Current status and future challenges. *Agronomy* **2018**, *8*, 57. [CrossRef]
4. Chen, J.; Fan, Y.; Wang, T.; Zhang, C.; Qiu, Z.; He, Y. Automatic Segmentation and Counting of Aphid Nymphs on Leaves Using Convolutional Neural Networks. *Agronomy* **2018**, *8*, 129. [CrossRef]
5. Svensgaard, J.; Roitsch, T.; Christensen, S. Development of a Mobile Multispectral Imaging Platform for Precise Field Phenotyping. *Agronomy* **2014**, *4*, 322–336. [CrossRef]
6. Zhang, J.; Naik, H.S.; Assefa, T.; Sarkar, S.; Reddy, R.C.; Singh, A.; Ganapathysubramanian, B.; Singh, A.K. Computer vision and machine learning for robust phenotyping in genome-wide studies. *Sci. Rep.* **2017**, *7*, 44048. [CrossRef]
7. Lee, U.; Chang, S.; Putra, G.A.; Kim, H.; Kim, D.H. An automated, high-throughput plant phenotyping system using machine learning-based plant segmentation and image analysis. *PLoS ONE* **2018**, *13*, e0196615. [CrossRef]
8. Li, L.; Zhang, Q.; Huang, D. A Review of Imaging Techniques for Plant Phenotyping. *Sensors* **2014**, *14*, 20078–20111. [CrossRef]
9. Navarro, P.J.; Pérez, F.; Weiss, J.; Egea-Cortines, M. Machine learning and computer vision system for phenotype data acquisition and analysis in plants. *Sensors* **2016**, *16*, 641. [CrossRef]
10. Tang, X.; Liu, M.; Zhao, H.; Tao, W. Leaf extraction from complicated background. In Proceedings of the 2009 2nd International Congress on Image and Signal Processing, Tianjin, China, 17–19 October 2009; pp. 1–5. [CrossRef]
11. Jiaofei, W.; Shuangxi, W.; Yanli, C. Research on the color image segmentation of plant disease in the greenhouse. In Proceedings of the 2011 International Conference on Consumer Electronics, Communications and Networks, Xianning, China, 16–18 April 2011; pp. 2551–2553.
12. Minervini, M.; Abdelsamea, M.; Tsaftaris, S.A. Image-based plant phenotyping with incremental learning and active contour. *Ecol. Inform.* **2014**, *23*, 35–48. [CrossRef]
13. Zhou, Q.; Wang, Z.; Zhao, W.; Chen, Y. Contour-based plant leaf image segmentation using visual saliency. In Proceedings of the International Conference on Image and Graphics, Tianjin, China, 13–16 August 2015; pp. 48–59.
14. Xu, L.; Lv, J. Recognition method for apple fruit based on SUSAN and PCNN. *Multimed. Tools Appl.* **2018**, *77*, 7205–7219. [CrossRef]
15. Scharr, H.; Minervini, M.; French, A.P.; Klukas, C.; Kramer, D.M.; Liu, X.; Luengo, I.; Pape, J.-M.; Polder, G.; Vukadino, D.; et al. Leaf segmentation in plant phenotyping: A collation study. *Mach. Vis. Appl.* **2016**, *27*, 585–606. [CrossRef]
16. Zhu, A.; Yang, L. An improved FCM algorithm for ripe fruit image segmentation. In Proceedings of the 2013 IEEE International Conference on Information and Automation (ICIA), Yinchuan, China, 26–28 August 2013. [CrossRef]
17. Pham, V.H.; Lee, B.R. An image segmentation approach for fruit defect detection using k-means clustering and graph-based algorithm. *Vietnam J. Comput. Sci.* **2015**, *2*, 25–33. [CrossRef]
18. Yamamoto, K.; Guo, W.; Yoshioka, Y.; Ninomiya, S. On plant detection of intact tomato fruits using image analysis and machine learning methods. *Sensors* **2014**, *14*, 12191–12206. [CrossRef]
19. Ubbens, J.; Cieslak, M.; Prusinkiewicz, P.; Stavness, I. The use of plant models in deep learning: An application to leaf counting in rosette plants. *Plant Methods* **2018**, *14*, 6. [CrossRef]
20. Deng, M.H.; Li, Z.C.; Zhu, S.P. The Agriculture Vision Image Segmentation Algorithm Based on Improved Quantum-Behaved Particle Swarm Optimization. *Appl. Mech. Mater.* **2015**, *713*, 1947–1950. [CrossRef]
21. Wu, N.; Li, M.; Chen, S.; Yuan, Y.; Zeng, X.; Chen, L.; Sun, X.; Bian, C. Automatic segmentation of plant disease images based on graph cuts fusing multiple features. *Trans. Chin. Soc. Agric. Eng.* **2014**, *30*, 212–219.

22. Shaikh, R.A.; Li, J.P.; Khan, A.; Khan, I. Content based grading of fresh fruits using Markov random field. In Proceedings of the 2016 3rd International Conference on Computing for Sustainable Global Development, New Delhi, India, 16–18 March 2016; pp. 3927–3931.
23. Hung, C.; Nieto, J.; Taylor, Z.; Underwood, J.; Sukkarieh, S. Orchard Fruit Segmentation using Multi-spectral Feature Learning. In Proceedings of the IEEE/RSJ International Conference on Intelligent Robots & Systems, Tokyo, Japan, 3–7 November 2013. [CrossRef]
24. Wang, X.; Ma, X.; Grimson, E. Unsupervised activity perception by hierarchical bayesian models. In Proceedings of the 2007 IEEE Conference on Computer Vision and Pattern Recognition (2007), Minneapolis, MN, USA, 17–24 June 2007; pp. 1–8.
25. Rasiwasia, N.; Vasconcelos, N. Latent dirichlet allocation models for image classification. *IEEE Trans. Pattern Anal. Mach. Intell.* **2013**, *11*, 2665–2679. [CrossRef]
26. Wang, X.; Grimson, E. Spatial latent dirichlet allocation. In *Advances in Neural Information Processing Systems 20, Proceedings of the Twenty-First Annual Conference on Neural Information Processing Systems, Vancouver, BC, Canada, 3–6 December 2007*; Neural Information Processing Systems (NIPS): Vancouver and Whistler, BC, Canada, 2008; pp. 1577–1584.
27. Niu, Z.; Hua, G.; Gao, X.; Tian, Q. Spatial-DiscLDA for visual recognition. In Proceedings of the IEEE Computer Vision and Pattern Recognition, Providence, RI, USA, 20–25 June 2011; pp. 1769–1776. [CrossRef]
28. Lacoste-Julien, S.; Sha, F.; Jordan, M.I. DiscLDA: Discriminative learning for dimensionality reduction and classification. In *Advances in Neural Information Processing Systems 21, Proceedings of the Twenty-Second Annual Conference on Neural Information Processing Systems, Vancouver, British Columbia, Canada, 8–11 December 2008*; Neural Information Processing Systems (NIPS): Vancouver and Whistler, BC, Canada, 2009; pp. 897–904. [CrossRef]
29. Ou, W.; Xie, Z.; Lv, Z. Spatially regularized latent topic model for simultaneous object discovery and segmentation. In Proceedings of the 2015 IEEE International Conference on Systems, Man, and Cybernetics, Kowloon, China, 9–12 October 2015; pp. 2938–2943. [CrossRef]
30. Li, Z.; Tian, W.; Li, Y.; Kuang, Z.; Liu, Y. A more effective method for image representation: Topic model based on latent dirichlet allocation. In Proceedings of the 2015 14th International Conference on Computer-Aided Design and Computer Graphics, Xi'an, China, 26–28 August 2015; pp. 143–148.
31. Niu, Z.; Hua, G.; Wang, L.; Gao, X. Knowledge-based topic model for unsupervised object discovery and localization. *IEEE Trans. Image Process.* **2018**, *27*, 50–63. [CrossRef]
32. Blei, D.M.; Ng, A.Y.; Jordan, M.I. Latent dirichlet allocation. *J. Mach. Learn. Res.* **2003**, *3*, 993–1022. [CrossRef]
33. Putthividhy, D.; Attias, H.T.; Nagarajan, S. Topic regression multi-modal latent dirichlet allocation for image annotation. In Proceedings of the 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, San Francisco, CA, USA, 13–18 June 2010; Volume 238, pp. 3408–3415. [CrossRef]
34. Wang, J.; Zhou, J.; Xu, H.; Mei, T.; Hua, X.S.; Li, S. Image tag refinement by regularized latent Dirichlet allocation. *Comput. Vis. Image Underst.* **2014**, *124*, 61–70. [CrossRef]
35. Yang, X.; Xu, D.; Qi, Y.J. Bag-of-words image representation based on classified vector quantization. In Proceedings of the 2010 International Conference on Machine Learning and Cybernetics, Qingdao, China, 11–14 July 2010; Volume 2, pp. 708–712.
36. Farhangi, M.; Soryani, M.; Fathy, M. Informative visual words construction to improve bag of words image representation. *IET Image Process.* **2014**, *8*, 310–318. [CrossRef]
37. Griffiths, T.L.; Steyvers, M. Finding scientific topics. *Proc. Natl. Acad. Sci. USA* **2004**, *101*, 5228–5235. [CrossRef]
38. Blei, D.; Carin, L.; Dunson, D. Probabilistic topic models. *IEEE Signal Process. Mag.* **2010**, *27*, 55–65. [CrossRef]
39. Noda, H.; Shirazi, M.N.; Kawaguchi, E. MRF-based texture segmentation using wavelet decomposed images. *Pattern Recognit.* **2002**, *35*, 771–782. [CrossRef]
40. Bharkad, S.; Kokare, M. Fingerprint matching using discreet wavelet packet transform. In Proceedings of the 2013 3rd IEEE International Advance Computing Conference (IACC), Ghaziabad, India, 22–23 February 2013; pp. 1183–1188. [CrossRef]
41. Vedaldi, A.; Soatto, S. Quick shift and kernel methods for mode seeking. In Proceedings of the European Conference on Computer Vision, Marseille, France, 12–18 October 2008; pp. 705–718. [CrossRef]
42. Comaniciu, D.; Meer, P. Mean shift: A robust approach toward feature space analysis. *IEEE Trans. Pattern Anal. Mach. Intell.* **2002**, *24*, 603–619. [CrossRef]

43. Sheikh, Y.A.; Khan, E.A.; Kanade, T. Mode-seeking by medoidshifts. In Proceedings of the 2007 IEEE 11th International Conference on Computer Vision, Rio de Janeiro, Brazil, 14–21 October 2007; pp. 1–8. [CrossRef]
44. Arias-Castro, E.; Mason, D.; Pelletier, B. On the estimation of the gradient lines of a density and the consistency of the mean-shift algorithm. *J. Mach. Learn. Res.* **2016**, *17*, 1487–1514.
45. Chacón, J.E. A population background for nonparametric density-based clustering. *Stat. Sci.* **2015**, *30*, 518–532. [CrossRef]
46. Genovese, C.R.; Perone-Pacifico, M.; Verdinelli, I.; Wasserman, L. Non-parametric inference for density modes. *J. R. Stat. Soc. Ser. B* **2016**, *78*, 99–126. [CrossRef]
47. Jiang, H.; Kpotufe, S. Modal-set estimation with an application to clustering. *arXiv*, 2016; arXiv:1606.04166.
48. Dasgupta, S.; Kpotufe, S. Optimal rates for k-nn density and mode estimation. In *Advances in Neural Information Processing Systems 27, Proceedings of the 28th Annual Conference on Neural Information Processing Systems 2014, Montreal, QC, Canada, 8–13 December 2014*; Neural Information Processing Systems (NIPS): Montreal, QC, Canada; pp. 2555–2563.
49. Winn, J.; Criminisi, A.; Minka, T. Object categorization by learned universal visual dictionary. In Proceedings of the Tenth IEEE International Conference on Computer Vision, Beijing, China, 17–21 October 2005; pp. 1800–1807. [CrossRef]
50. Bertolino, P.; Montanvert, A. Multiresolution segmentation using the irregular pyramid. In Proceedings of the 3rd IEEE International Conference on Image Processing, Lausanne, Switzerland, 19 September 1996; Volume 1, Volume 257–260. [CrossRef]
51. Sumengen, B.; Manjunath, B.S. Multi-scale edge detection and image segmentation. In Proceedings of the 2005 13th European Signal Processing Conference, Antalya, Turkey, 4–8 September 2005; pp. 1–4.
52. Yang, Y.; Xu, L. Remote sensing image classification using layer-by-layer feature associative conditional random field. *J. Comput. Appl.* **2014**, *34*, 1741–1745. [CrossRef]
53. Li, H.; Meng, F.; Wu, Q.; Luo, B. Unsupervised Multiclass Region Cosegmentation via Ensemble Clustering and Energy Minimization. *IEEE Tran. Circuits Syst. Video Technol.* **2014**, *24*, 789–801. [CrossRef]