Journal of Biomedical Informatics 44 (2011) 848–858

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

Automatic figure classification in bioscience literature

Daehyun Kim *, Balaji Polepalli Ramesh, Hong Yu

Department of Health Sciences, University of Wisconsin-Milwaukee, Milwaukee, WI, USA

ARTICLE INFO

Article history:
Received 17 September 2010
Accepted 11 May 2011
Available online 27 May 2011

Keywords:
Classification
Taxonomy
Hierarchical model
Machine learning model

ABSTRACT

Millions of figures appear in biomedical articles, and it is important to develop an intelligent figure search engine to return relevant figures based on user entries. In this study we report a figure classifier that automatically classifies biomedical figures into five predefined figure types: *Gel-image*, *Image-of-thing*, *Graph*, *Model*, and *Mix*. The classifier explored rich image features and integrated them with text features. We performed feature selection and explored different classification models, including a rule-based figure classifier, a supervised machine-learning classifier, and a multi-model classifier, the latter of which integrated the first two classifiers. Our results show that feature selection improved figure classification and the novel image features we explored were the best among image features that we have examined. Our results also show that integrating text and image features achieved better performance than using either of them individually. The best system is a multi-model classifier which combines the rule-based hierarchical classifier and a support vector machine (SVM) based classifier, achieving a 76.7% F1-score for five-type classification. We demonstrated our system at http://figureclassification.askhermes.org/.

1. Introduction

A picture is worth a thousand words. In bioscience, figures are usually a part of the evidence of biomedical experiments, and biomedical researchers incorporate numerous figures into their publications to report experimental results, present research models, and provide examples of biomedical objects (e.g., cells, tissue, and organs). For example, a random collection of 1750 biological articles in *Proceedings of the National Academy of Science* (PNAS) in 2010 show an average of 4.2 figures. Physicians may want to access biomedical images reported in the literature for the purpose of clinical education or to assist clinical diagnoses. Additionally, biologists may want to identify images that support specific biological phenomena or experiment results. Therefore, we are developing an approach for classifying figures into figure types as a potential way to facilitate figure searching (http://figuresearch.askhermes.org) in the biomedical literature.

Bioscience figure mining has become a research trend only recently, and most of the research in this area has been text-centric. For example, Yu and Lee [1] found a connection between sentences appearing in the abstract and figures appearing in the same article, such that the biological figures could be accessed by a link in abstract sentences. Furthermore, Yu et al. [2] evaluated associated text for figure comprehension and then developed summarization technologies to automatically generate a structured summary for each figure [3]. The evaluation study has shown that such a summary automatically generated is useful for figure comprehension [3]. Yu et al. [4] developed natural language processing (NLP) approaches to rank figures based on their biological importance and evaluated novel user interfaces that incorporate figure ranking. BioText [5], GoldMiner [6], BioMed Search [7], and Yale Image Finder (YIF) [8] applied figure captions or text appearing in a figure for figure searching.

In addition to text associated with figures, figures themselves are important for figure mining. For instance, when a physician searches for "lung cancer" figures from the literature for the purpose of diagnosis, he or she might prefer finding a microscopic lung cancer figure rather than a lung cancer statistics graph or chart. Most of the systems described above [1,5–8] do not currently incorporate such figure classification, such as the YIF system shown in Fig. 1.

In general, biomedical image classification systems have used both text and image features. For instance, the SLIF (Subcellular Location Image Finder) system used image features simultaneously with text features for subfigure detection [9,10], figure topic detection [11], and figure search [12,13]. In addition, Shatkay et al. [14] showed that derived image features improve document categorization. To incorporate image features for document classification, figures were first segmented into subfigures and subsequently each subfigure was classified into one of seven predefined figure types. Figure types of all subfigures were then added as learning features for document classification. Previously, we integrated an image feature-based figure classifier and a text feature-based

* Corresponding author. Address: 2400 E Hartford Ave., Milwaukee, WI 53211, USA. Fax: +1 414 229 2900.

E-mail addresses: kim48@uwm.edu (D. Kim), brpjr@uwm.edu (B.P. Ramesh), hongyu@uwm.edu (H. Yu).

1532-0464/$ - see front matter © 2011 Elsevier Inc. All rights reserved.
doi:10.1016/j.jbi.2011.05.003

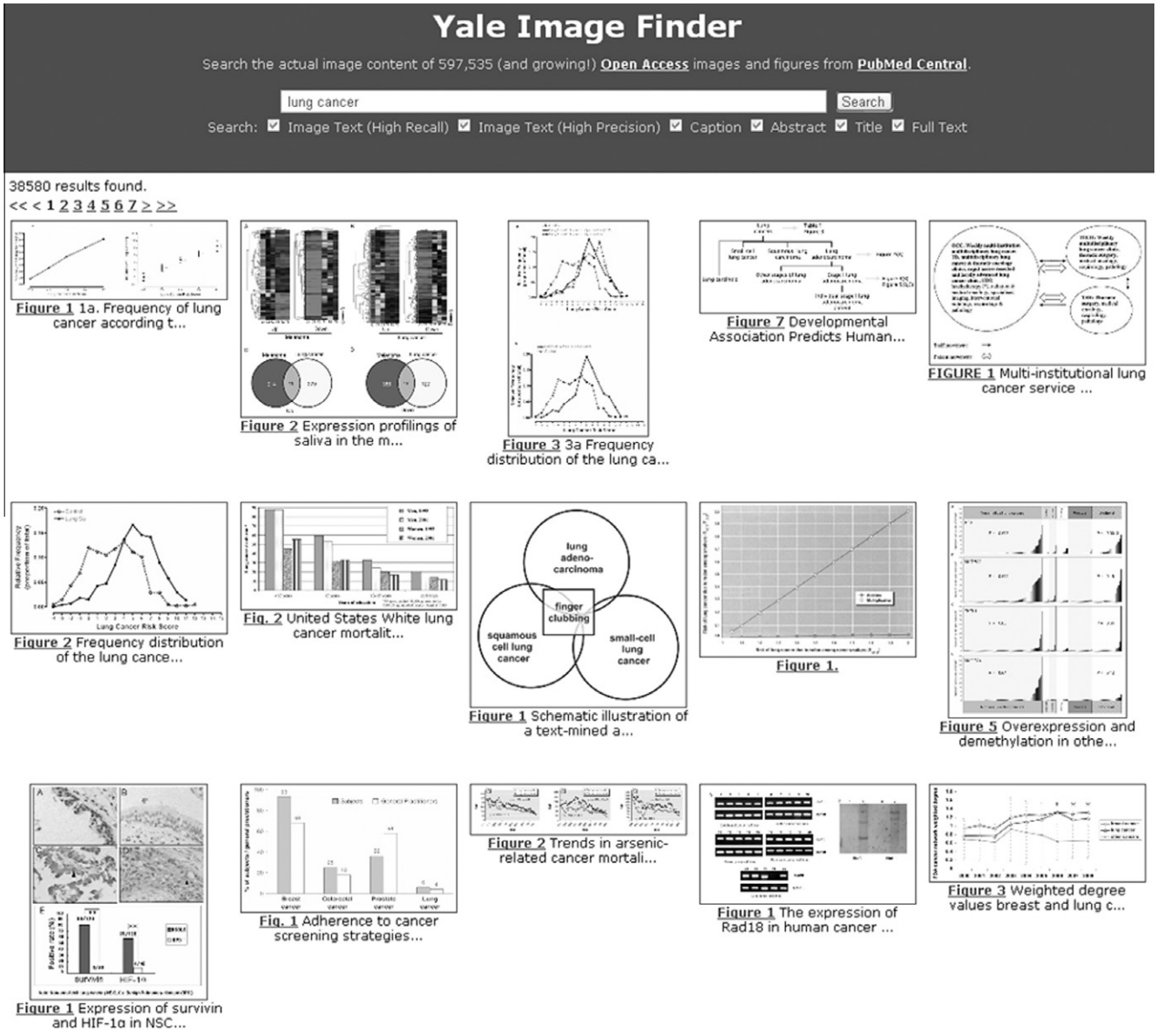

Fig. 1. Illustration of YIF Figure Search System with query of "lung cancer". An illustration of existing figure search engine for a query "lung cancer".

figure classifier and the results showed that the integration improved figure classification [15].

Existing approaches in figure classification were limited in that they did not explore image feature analysis, nor did they combine different strengths from different classification models, as we did in this study [9–15]. As reported in our previous work [15], we manually examined images to create a generic high-level image taxonomy (example figures are shown in Fig. 2), as shown below:

- *Gel-image* consists of gel images including Northern (for DNA), Southern (for RNA), and Western (for protein).
- *Image-of-thing* refers to images and photographs of existing objects such as cells (microscopic), cell components, tissues, organs, species, and photographs of laboratory equipment.
- *Graph* consists of bar charts, column charts, line charts, plots, and other graphs that are drawn either by authors or by a computer.
- *Model* may demonstrate a biological process, molecular docking, or an experimental design. We include any structure (e.g., chemical, molecular, or cellular) as well as gene or protein sequences and their alignments as *Model*.
- *Mix* refers to a figure that incorporates two or more other figure types.

The above taxonomy attempts to represent a standard image classification (e.g., *Graph* and *Image-of-thing*) as well as being biologically knowledge-driven (e.g., *Gel-image* and *Model*). Such taxonomy can be further refined with children categories, as shown in [14].

We hypothesize that certain image features strongly associate with specific figure types, and such 'surrogate' image features can be used to improve figure classification. For instance, as shown in Fig. 3, of the four commonly used intensity histogram features of images (i.e., mean, variance, skew, and entropy) [9,10,14,15], only skew was strongly associated with a single figure type, as it presented a positive value for *Image-of-thing* and negative values for the other figure types. This is because *Image-of-thing* usually does not have white background space as much as other figure types, and thus its intensity histogram mostly incorporates low grey-level, which results in a positive skew value. Other features

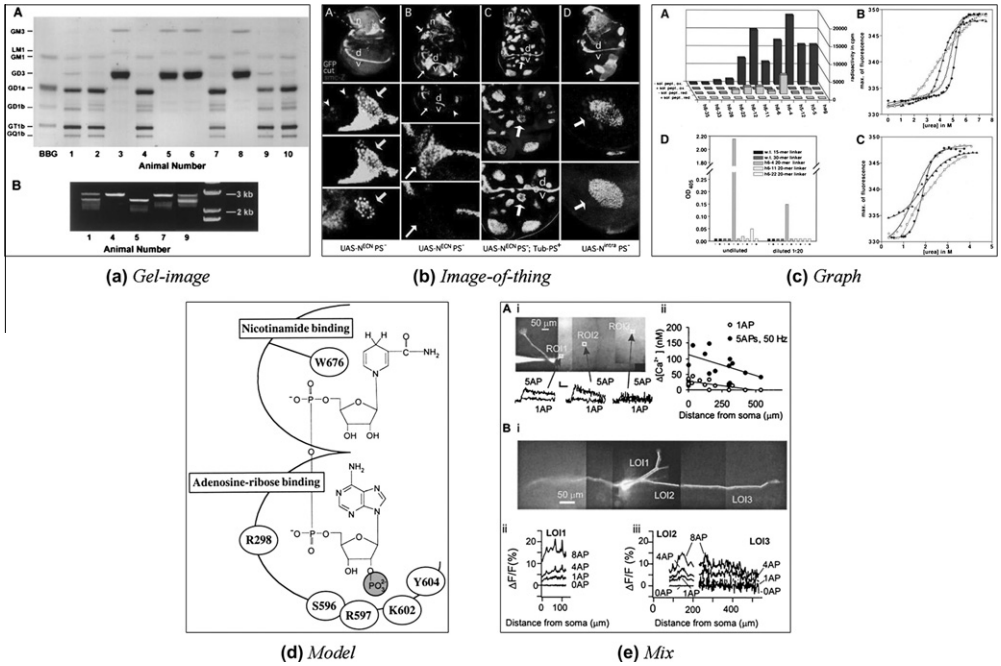

Fig. 2. The five figure types of biomedical figures. An illustration of five figure types of biomedical figures in the publications.

(mean, variance, and entropy) do not strongly associated with figure types. In this study, we identified distinctive image features (e.g., skew and uniformity) for each figure type and evaluated them for figure classification. We also explored new image features not yet explored in any of the previous studies [9,10,14,15].

In additional to image features, we also explored different models for figure classification. In our previous work [15], we built two different supervised machine learning classifiers that were trained on text features from a caption and image features from a figure, respectively, and then combined the results of the two classifiers. In contrast, in this study, we first report a single supervised machine-learning classifier that was trained on both image and text features. We then report a rule-based hierarchical figure classifier which learns from figure-type-specific image features. Finally, we report a multi-model classifier that integrates the supervised machine-learning classifier and the rule-based hierarchical classifier.

2. Methods

2.1. Data and annotation agreement

This study followed the same figure-type taxonomy established in our earlier work [15], and reused the same 767 annotated figures [15]. As reported previously [15], the 767 figures were randomly selected from a pool of 88,225 biomedical figures appearing in biological articles published in PNAS (year 1995–

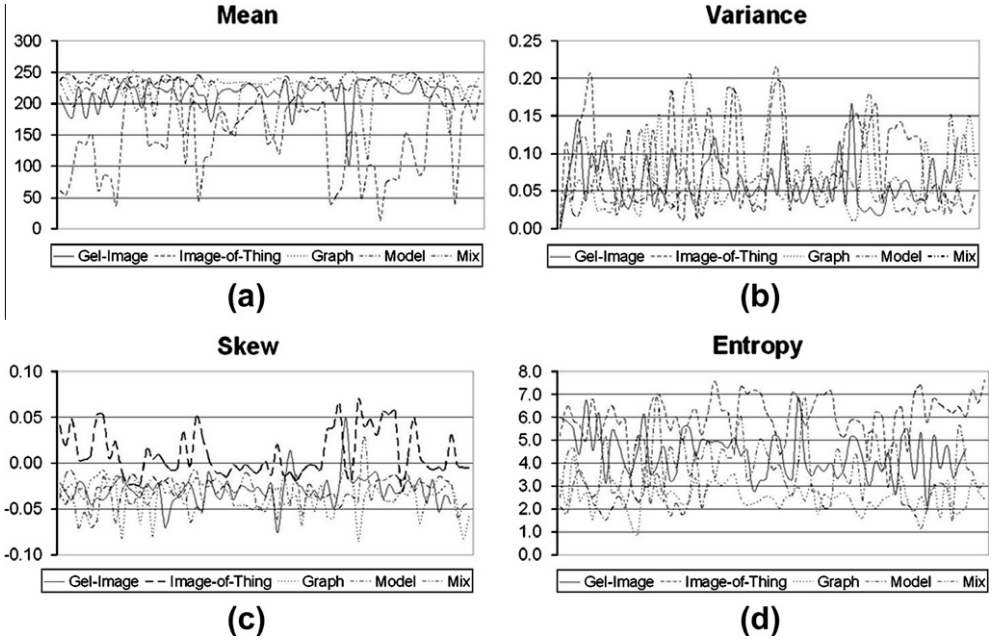

Fig. 3. Examples of image features according to figure types. An illustration of distributions of image features according to figure types.

2005). Each figure was judged by the last author (HY) to be one of the following five figure types: *Gel-image* (119), *Image-of-thing* (90), *Graph* (264), *Model* (152), and *Mix* (142) [15]. No inter-annotator agreement, however, was reported.

We recruited a biologist (BS in biology) who is not the author of this manuscript. Fifty figures were randomly selected from the 767 annotated figures. Biologists assigned independently categories defined in [15] to each figure. We then measure the inter-annotator agreement between the biologist and HY and report the annotation agreement with the Cohen's kappa value [16] and the overall agreement.

2.2. Feature representation

We explored three feature types: image feature, text feature, and joint feature, the latter of which integrates both image and text features. For image features, we explored both common image features as well as new image features we derived from biomedical figures.

2.2.1. Common image features

We explored histograms, which are graphical representations of the tonal distribution in an image [17] that were commonly used for image classification [14,15,18,19]. Specifically, we explored colour histograms (CH), intensity histograms (IH), and edge direction histograms (EDH). The results (Fig. 6c) show that only *Image-of-thing* incorporates colour, while most of other figure types exhibit grey-level distribution. As a result, we did not select CH as an image feature for our figure type classification.

IH represents the distribution of pixels in a figure according to their grey-level presentation. We explored five statistical values from IH: the first three moments (mean, variance, and skew), entropy, and uniformity [17]. EDH is typically used for shape-based retrieval [20–22]; we computed EDH by grouping the edge pixels that fall into edge directions, and then counting the number of pixels in each direction. We detected edges using the Sobel operator [17] and used a bin granularity of 1°, resulting in a histogram of 180 bins. We also explored the same five statistical values from EDH as we had done for IH.

2.2.2. New image features

In addition to histograms, we explored new image features derived from biomedical figures, including set of image features (e.g., moments, entropy, and uniformity) evaluated previously [15,19] and new image features derived from subfigures.

Biomedical figures typically incorporate multiple subfigure figures (one example is shown in Fig. 2). Our training data (details in Section 2.4) shows that 84.9% of the figures incorporated subfigures. *Mix* must incorporate multiple subfigures; however, a figure that incorporates subfigures is not necessarily a *Mix* figure, Fig. 2a–d are such examples. We derive a set of new image features from those subfigures. This is done by segmenting figure into its subfigures – called subfigure segmentation – and then computing image features from the subfigures.

2.2.2.1. Subfigure segmentation. Subfigure segmentation has been studied in the biomedical literature. Shatkay et al. [14] segmented subfigures based on connected component analysis [17]. Murphy et al. [9] detected image boundaries through recursive panel splitting. These approaches, however, either lacked evaluation or were developed for certain figure type: focusing on black-and-white images only [14] or fluorescent microscope images [9]. Kou et al. [23] used subfigure labels to segment subfigures. However, we have found that often there are two or more subfigures under one subfigure label, an example of which is shown in Fig. 2e.

We therefore developed new approaches for biomedical subfigure segmentation in the context of figure classification. Since the boundaries of subfigures in *Gel-image* and *Image-of-thing* were relatively easier to detect and those of *Graph* and *Model* posed challenges, we developed a two-step approach to segment *Gel-image* and *Image-of-thing*.

At the first step, we separated *Graph* and *Model* from *Gel-image* and *Image-of-thing*. This was done by applying Fourier analysis [17], which showed that *Gel-image* and *Image-of-thing* primarily comprised low-frequency signals, while *Graph*, *Model*, and textual descriptions were composed of high-frequency signals, as shown in Fig. 4. We used the Sobel edge operator to separate *Graph*, *Model*, and textual description from *Gel-image* and *Image-of-thing*, a process that is shown in Fig. 5.

Specifically, we first applied the Sobel operator to detect high-frequency regions (textual description in Fig. 5a). Note that the boundary of gel image was also detected (the rectangular white lines as shown in Fig. 5b) because they have high-frequency signals. To reserve the boundary of low-frequency regions (gel image as shown in Fig. 5a), we applied morphological techniques to refine the detected regions (Fig. 5c) [17] and then removed high-frequency regions from input figures and assigned them the same background colour (Fig. 5d).

At the second step, we applied the commonly used connected component analysis to segment *Gel-image* and *Image-of-thing* and then evaluated the segmentation results with the training data that incorporated a total of 416 *Graph* and *Model* subfigures and 952 *Gel-image* and *Image-of-thing* subfigures. Our evaluation results show that this approach successfully removed 96% of the *Graph* and *Model* subfigures and correctly segmented 89% of the *Gel-image* and *Image-of-thing* subfigures.

2.2.2.2. Heuristic image feature extraction. Once subfigures were segmented, we constructed IHs from an input figure (F_I), its segmented subfigures (F_S) (e.g., Fig. 5d), and individual segmented subfigures (F_i). Image features were then evaluated for their statistical values (e.g., average and variance) so that distinctive image features can be identified. As shown in Fig. 6, we found eight heuristic image features (HIF) for figure classification:

- *Skew difference:* For F_I, as shown in Fig. 6a, the skew of *Model* is higher but less variable than *Graph*. Therefore, we can limit the range of *Model* using the following formula: $|\gamma - \bar{\gamma}_{Model}|$, where γ is the skew value of F_I and $\bar{\gamma}_{Model}$ is the average skew value of the training data in *Model*. If the absolute difference is less than a certain threshold, it may belong to *Model*.
- *Variance difference:* In contrast to skew, the variance of *Model* is lower and less variable than *Graph*, as shown in Fig. 6b. Accordingly, we can also limit the range of *Model* with the following formula: $|\sigma - \bar{\sigma}_{Model}|$, where σ is the variance value of F_I and $\bar{\sigma}_{Model}$ is the average variance value of the training data in *Model*. If the absolute difference is less than a certain threshold, it may belong to *Model*, as well.
- *Colour density: Gel-image* contains almost no colour figures, while *Image-of-thing* contains numerous colour figures. Therefore, we assumed the colour density to be a useful feature for separating *Gel-image* from *Image-of-thing*. Specifically, we counted the number of colour pixels in F_S and computed the colour density as (the number of colour pixels)/(the total number of pixels in F_S). We then set it as colour figure if the colour density is higher than 10%. As shown in Fig. 6c,
- *Subfigure uniformity:* Fig. 6d shows that the uniformity values of F_S in *Graph* and *Model* are greater than in other figure types because most of the image content became a constant (i.e., white) background in subfigure segmentation.

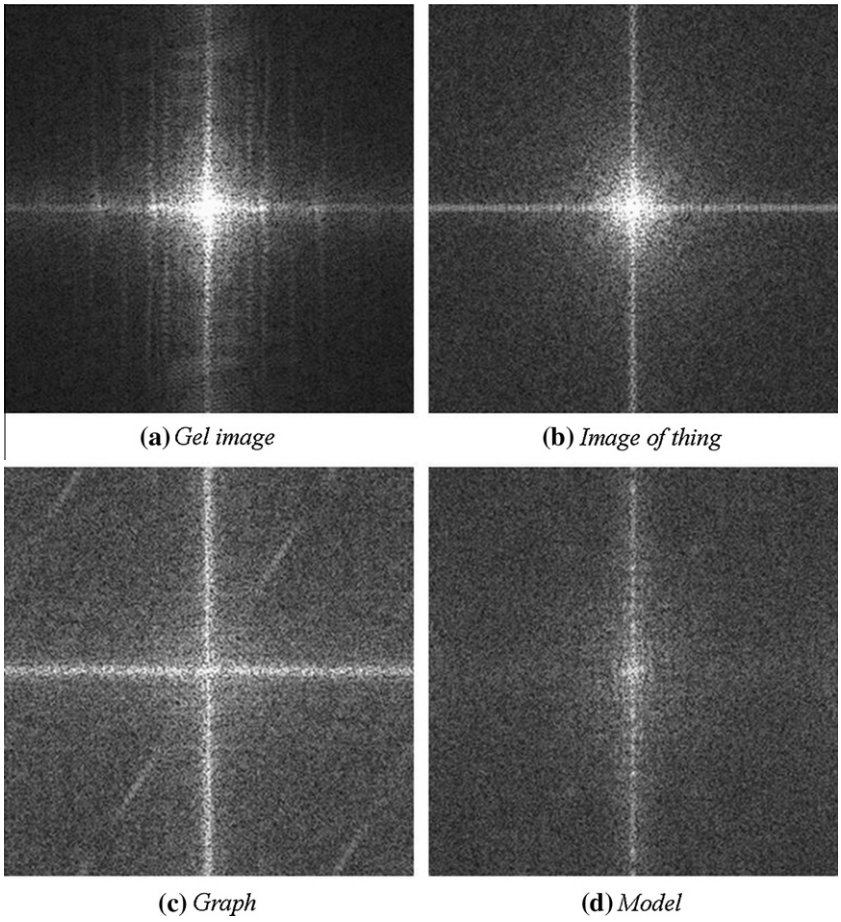

Fig. 4. Frequency spectrums of figure types. Most frequencies of both *Gel-image* and *Image-of-thing* are existed around centre of spectrum images (i.e., low frequency); while both *Graph* and *Model* almost have high frequencies.

- *Entropy ratio:* We computed the ratio of entropy values between F_I and F_S as (entropy of F_S)/(entropy of F_I). As shown in Fig. 6e, *Image-of-thing* undergoes a very small change in entropy ratio, while other figure types experienced more significant changes, particularly *Model*. Thus, a small entropy ratio can be used for classifying *Image-of-thing* and *Model*.
- *Maximum uniformity difference:* We first computed the uniformity values of F_i, and then we found the minimum and maximum uniformity values. Hereby, the maximum uniformity difference was measured by subtracting the minimum uniformity value from the maximum uniformity value. Since subfigures (F_i) of *Mix* present different uniformity values due to different subfigure types, the maximum uniformity difference of *Mix* subfigures should be much higher than other figure types. Fig. 6f illustrated an example. Thus, maximum uniformity difference can be used as a specific feature of *Mix*.
- *Residual pixels:* A residual pixel is a segmented pixel in the subfigure segmentation. The number of residual pixels of F_S in *Graph* and *Model* is extremely small because they lose most of their image content during subfigure segmentation. We can compute the ratio of residual pixels between F_I and F_S using the following formula:

$$r = \frac{\text{number of residual pixels in } F_S}{\text{number of residual pixels in } F_I}$$

As shown in Fig. 6g, the ratio of residual pixels of *Gel-image* and *Image-of-thing* are high, while the ratios of *Graph* and *Model* are very low. As a result, the ratio of residual pixels can be used to discriminate *Gel-image* and *Image-of-thing* from other figure types.

- Distance between centres of gravity: In general, the centre of gravity of F_S (C_S) is very similar to the centre of gravity of F_I (C_I), as illustrated in Fig. 7a. However, if one of the subfigures belongs to *Graph* and *Model* and disappears after subfigure segmentation, then C_S may move far from C_I, as shown in Fig. 7b. Fig. 6h shows that for *Gel-image* and *Image-of-thing*, the distance between C_I and C_S is very small. In the case that all subfigures disappear, C_S moves to the left-top corner of the figure and the distance increases. For *Mix*, the distance is intermediate. Thus, the distance between C_I and C_S is useful for classifying figure types.

2.2.3. Text features

Previous work concluded that text features are crucial for image classification [1,5–7,14,15,24] and we therefore explored text features for figure classification. We first extracted word features from the captions of each figure and then processed them by excluding numbers and special symbols (e.g., @, #, ∗, %, etc.), removing stop words [25], and applying the Porter stemming [26]. A χ^2 test [27] was then applied to determine which textual terms had a significant correlation with specific figure types. Using a cutoff of $P < 0.05$ resulted in 568 text features, e.g., "land," "gel," and "protein" for *Gel-image*; "stain," "microscopy," and "phenotype" for *Image-of-thing*; "response," "sequence," and "measure" for *Graph*; "box," "cell," and "model" for *Model*; and "fragment," "blot," and "size" for *Mix*. In this work, we did not consider any extra process to deal with synonyms separately, but it might be worth considering synonyms for future versions.

A textual term vector was used for representing the text features for each figure I_i. Similar to the vector space model [28–30] used in

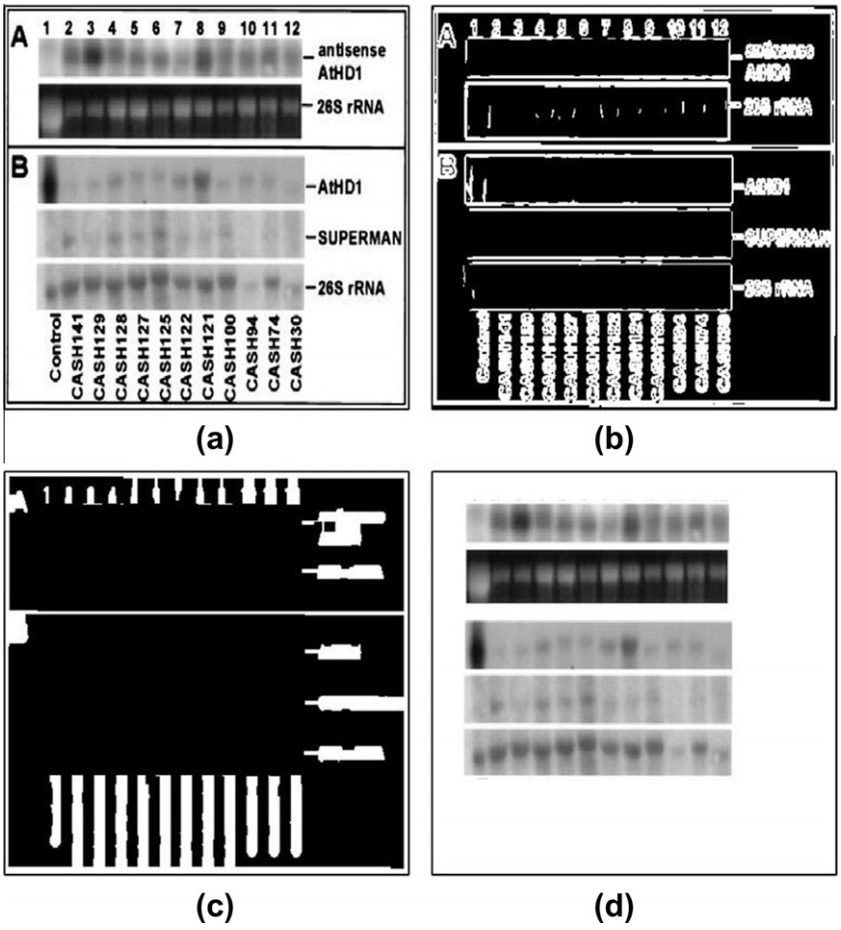

Fig. 5. Subfigure segmentation. Proposed subfigure segmentation method consists of four major steps: (a) whole figure, (b) detected high-frequency signals, (c) refined high-frequency signals, (d) segmented subfigures.

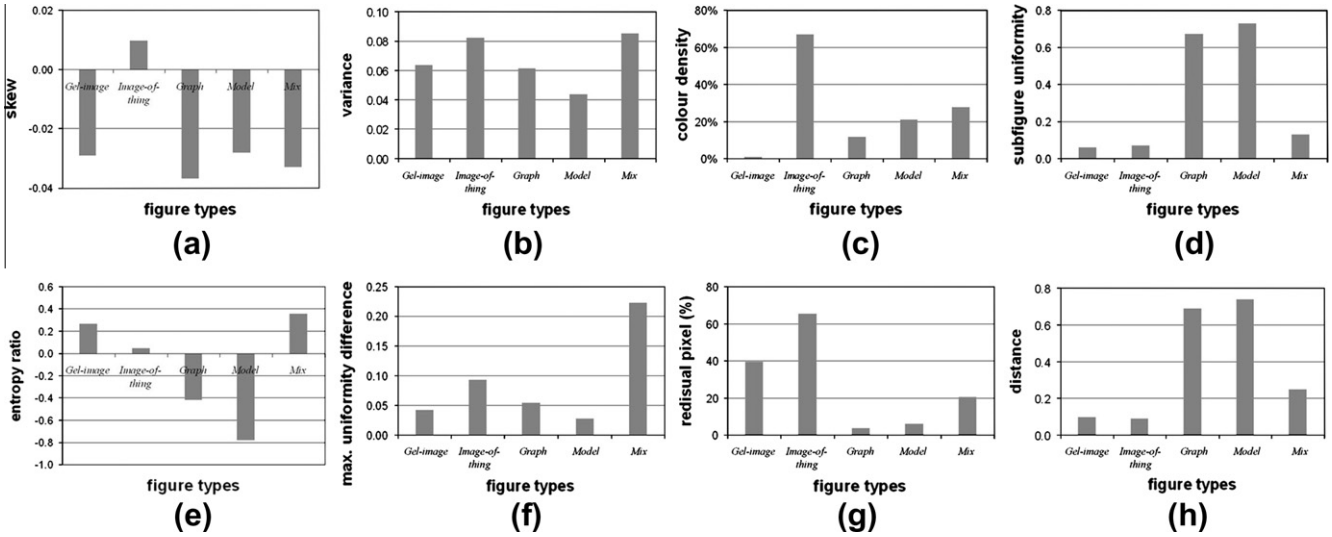

Fig. 6. Attributes of image features. An illustration of distributions of image feature values (i.e., average and variance).

information retrieval, each term can also be weighted by TFIIF (term frequency inverse image frequency) [18]. TFIIF can be defined as

$$f_i^{(T)} = [t_{i,1} \cdot \log(N/n_1), \ldots, t_{i,j} \cdot \log(N/n_j), \ldots, t_{i,m} \cdot \log(N/n_m)],$$

where $t_{i,j}$ is the frequency of term j appearing in the figure caption of figure I_i; n_j is the number of figures characterized by term j; and m and N are the number of lexical cues and the total number of figures in the training data set, respectively.

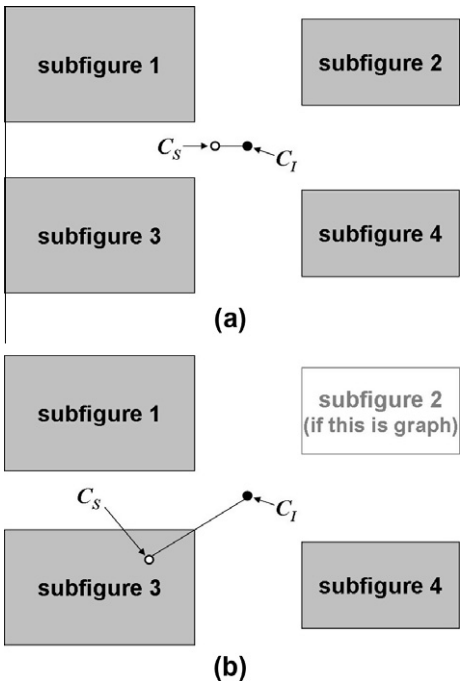

Fig. 7. Average distance between a figure's centre and the centre of its subfigures. An illustration of average distance between a figure's centre and the centre of its subfigures: (a) before subfigure segmentation, (b) after subfigure segmentation.

2.2.4. Joint features

Image features ($f^{(I)}$) and text features ($f^{(T)}$) that we described in Sections 2.2.1–2.2.3 can be combined into a high-dimensional vector for figure classification. In this study, we adopted the models of Tian et al. [18] and Cascia et al. [31] to integrate image and text features into a high-dimensional joint feature vector like $f^{(J)} = [f^{(I)}, f^{(T)}]$. Such models allow us to develop a simple single-figure classifier. Because image and text features can have quite different variations, we normalized each feature vector in the joint space according to its maximum elements [7,32].

2.3. Figure classifiers

We explored three different classifiers for figure classification: a rule-based hierarchical figure classifier, a supervised machine-learning classifier, and a multi-model figure classifier that combines both.

2.3.1. Rule-based figure classifier

As described in Section 2.2.2.1, we found that *Gel-image* and *Image-of-thing* mostly incorporated low-frequency signals, while *Graph* and *Model* were composed of high-frequency signals. On the basis of this distinction, *Gel-image* and *Image-of-thing* were grouped as a texture group and *Graph* and *Model* as a nontexture group. *Mix* can potentially belong to either group based on the composition of its subfigures. Accordingly, we developed a rule-based figure classifier that first classified a figure into either texture or nontexture group and then further classified it into its figure type. The process is shown in Fig. 8.

We found that *Graph* and *Model* lose nearly all high-frequency signals after subfigure segmentation. As a result, the image features – the uniformity and residual pixels of F_S (Fig. 6d and g, respectively) – can be used to distinguish *Gel-image* and *Image-of-thing* from *Graph* and *Model*. Moreover, as we mentioned earlier, skew is strongly associated with *Image-of-thing*. Therefore, we used these three image features for classifying figures into texture and nontexture groups. Fig. 9a shows the pseudocode at the top layer of our rule-based hierarchical figure classifier.

At the bottom layer, we explored nine image features, as shown in Table 2. As described in Section 2.2.2.2, we explored the image features ③ and ①⓪ for *Gel-image*; ①, ②, ⑥, and ⑦ for *Image-of-thing*; ⑤ for *Mix*; and ⑧ and ⑨ for *Model*. Fig. 9b and c show the pseudocode for further classifying a figure into its figure type in each group. All threshold values (t_1–t_{12}) in the pseudocode were empirically determined based on statistics information of the data excluding testing set.

2.3.2. Supervised machine-learning classifiers

As an alternative, we explored supervised machine-learning algorithms, including artificial neural networks (ANN), *K*-nearest neighbourhood (KNN), Naïve Bayes model (NB), and support vector machines (SVM). For SVM, we explored different kernel functions (e.g., linear, polynomial, radial basis function (RBF), and sigmoid). We fixed the order of polynomial kernel function at 2 and performed the grid search over the adjustable parameters for values of C and γ equal to $[10^{-5}, 10^{-4}, 10^{-3}, \ldots, 10^{5}]$. The ANN model constructed 10 neurons in the hidden layer and five in the output layer. The supervised machine-learning classifiers were implemented with OpenCV [33], which is an open-source computer vision library originally developed by Intel and commonly used in image processing and pattern recognition applications.

2.3.3. Multi-model figure classifier

We developed a multi-model figure classifier that integrates the two aforementioned classifiers. The multi-model figure classifier first applied the rule-based classifier to separate figures into texture and nontexture groups, and then applied the rule-based classifier to further separate the texture group *Gel-image* and *Image-of-thing*. In contrast, the nontexture group *Graph* and *Model* was separated by the supervised machine-learning classifier (specifically, SVM).

2.4. Evaluation methods

As described in Section 2.1, a total of 767 annotated figures were used as training and testing this study. Specifically, we used 50% of the data for training and the remaining 50% for testing. We performed 10-fold cross-validation and grid search [32,33]. The training data was also used to select image and text features and to determine threshold values of the hierarchical figure classifier and of certain image feature values (e.g., average skew and variance values as described in Section 2.2.2.2).

We evaluated the performance of our figure classifiers on the testing data, reporting recall, precision, and F1-score. Recall is the number of true positives divided by the total number of figures

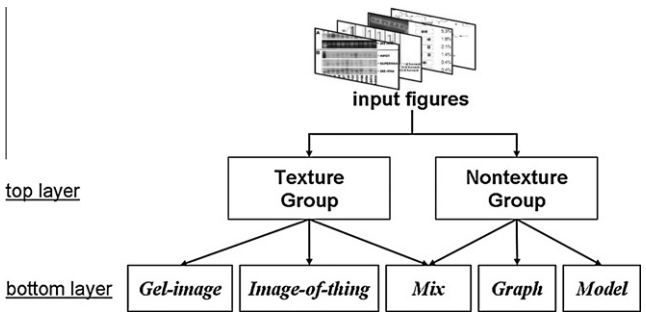

Fig. 8. Overall procedure of the hierarchical figure classifier. An illustration of the overall procedure of the hierarchical figure classifier.

If ② is positive,
    then classify it into the texture group
else if ④ is less than t_1 and ⑦ is more than t_2,
    then classify it into the texture group
else classify it into the nontexture group

(a) Pseudocode at the top layer

If ② is more than t_3,
    then classify it into *Image-of-Thing*
else if ① is true, ⑥ is less than t_4, and ⑦ is more than t_5,
    then classify it into *Image-of-Thing*
else if ⑤ is more than t_6,
    then classify it into *Mix*
else if ③ less than t_7 and ⑩ is less than t_8,
    then classify it into *Gel-Image*
else classify it into *Mix*

(b) Pseudocode of the texture group at the bottom layer

If ⑤ is more than t_9 and ⑦ is more than t_{10},
    then classify it into *Mix*
else if ⑧ is less than t_{11} and ⑨ is less than t_{12},
    then classify it into *Model*
else classify it into *Graph*

(c) Pseudocode of the nontexture group at the bottom layer

Fig. 9. Pseudocode of the rule-based figure classifier. An illustration of pseudocode of the hierarchical figure classifier.

of this type; precision is the number of true positives divided by the total number of figures recognized to be of this type; and the F1-score is the harmonic mean of recall and precision. To evaluate the overall performance of our figure classifiers, we used accuracy, which is the total number of true positives divided by the total number of figures.

3. Results

3.1. Annotation agreement

Of the 50 figures annotated by the biologist, 48 were consistent with our previous annotation. The results showed an agreement of 96% and a kappa value of 0.95 with 95% confidence.

3.2. Image feature analysis

Table 1 shows the accuracy of the four supervised machine-learning classifiers (ANN, KNN, NB, and SVMs) that were trained on seven image feature combinations, which included IH, EDH, and HIF described in Section 2.2.2.2. For SVM, we explored different kernel functions (linear, polynomial, RBF, and sigmoid). As shown in Table 1, among different feature combinations, RBF based SVM attained the best accuracy of 64.5% when using IH + HIF. Excluding mean, variance, and entropy of IH, the image features that do not associate strongly with specific figure types (as shown in Fig. 3), IH* + HIF achieved further performance improvement with the accuracy of 68.4%. Although the performance is not excellent, these image features are the best among image features that we have checked.

Based on the optimized feature combination IH* + HIF, Table 2 shows the corresponding 10 individual image features, which include eight heuristic image features (HIF) described in Section 2.2.2.2 and two image features of IH (i.e., skew and uniformity).

3.3. Performance comparison between image features and text features

Table 3 shows F1-scores of the four supervised machine-learning classifiers (i.e., ANN, KNN, NB, and SVMs) for three different feature types. The results indicate that different classifiers performed differently in terms of favouring text features, image features and their combination (joint). The results show that overall SVMs with different kernel functions outperformed all the other three classifiers. With image features alone, the SVM RBF attained the best performance of 65% accuracy and 60% F1-score, while the SVM linear outperformed other kernel functions on text features, attaining 71.0% accuracy and 69.6% F1-score, a significant improvement over using the image features alone. The best system was SVM polynomial that was trained on the joint text and image features, attaining 74.4% accuracy and 72.8% F1-score.

Table 4 shows that for *Gel-image* and *Image-of-thing*, the image-feature based SVM classifier performed better than the text-feature based SVM classifier, attaining F1-scores of 73.4% and 82.5%, respectively. In contrast, for *Graph* and *Model*, the text-feature based SVM classifier performed better than the image-feature based SVM classifier, attaining F1-scores of 78% and 78.4%, respectively. The joint image-text based SVM classifier performed the best for three figure types: *Image-of-thing*, *Graph*, and *Mix*, attaining F1-scores of 82.9%, 82.1%, and 52.5%, respectively.

3.4. Performance on three classification models

We evaluated the results of three classification models: the hierarchical figure classifier, the supervised machine-learning classifier, and the multi-model classifier. For the machine-learning system, we used the best SVM classifier with the joint feature setting.

Table 5 shows the F1-score, recall, and precision of three classifiers for five-way figure-type classification. The results indicate that the hierarchical classifier performed better on *Gel-image* and *Image-of-thing* than the SVM, attaining F1-scores of 75.6% and 86%, respectively. In contrast, the SVM performed better on *Graph* and *Model* than the hierarchical classifier, attaining F1-scores of 82.1% and 75.3%, respectively. That is, the SVM is more optimal to classify *Graph* and *Model* than the hierarchical classifier. Therefore, we applied the SVM for the nontexture group as described in Section 2.3.3, and as a result, the multi-model classifier outperformed both hierarchical and SVM classifiers for all figure types except for *Model*.

Table 6 shows the overall accuracy and F1-score for five-figure-type classification. The results show that the hierarchical figure classifier attained 62.7% accuracy and a 63.3% F1-score, which is 28.3% and 12.1% higher, respectively, than the baseline system, which assigns a figure to its highest frequently occurred type (*Graph*). The SVM classifier attained 74.4% accuracy and a 72.8% F1-score, which is 11.7% and 9.5% higher, respectively, than the rule-based hierarchical figure classifier. The multi-model figure classifier outperformed both, attaining 77.8% accuracy and a 76.7% F1-score.

Fig. 10 shows the variance of precision, recall, and F1-score for each classification model over all figure categories as in Table 5. We found that the rule-based hierarchical figure classifier had the largest variance of the classification performance, while the

Table 1
Accuracy comparison among different image feature type combinations.

Features	No. of features	ANN (%)	KNN (%)	NB (%)	SVM (RBF) (%)
IH	5	50.0	55.3	55.3	57.9
EDH	5	31.6	47.4	57.9	36.8
HIF	8	47.8	58.8	57.9	59.8
IH + EDH	10	56.4	62.9	62.4	34.5
IH + HIF	13	55.3	56.9	59.8	64.5
EDH + HIF	13	55.1	57.4	50.1	43.6
IH + EDH + HIF	18	61.6	60.6	61.6	35.5
IH* + HIF	10	60.5	63.2	60.5	68.4

IH: intensity histogram (e.g., mean, variance, skew, entropy, and uniformity).
EDH: edge direction histogram (e.g., mean, variance, skew, entropy, and uniformity).
HIF: heuristic image features (eight image features described in Section 2.2.2.2).
IH*: modified intensity histogram (i.e., mean, variance, and entropy of IH were excluded).
KNN: *K*-nearest neighbour, NB: Naïve Bayesian.
SVM: support vector machine, ANN: artificial neural network.

Table 2
Distinctive 10 image features including eight heuristic image features (HIF) and modified intensity histogram (IH*: skew and uniformity).

No.	Image features
①	Colour density
②	Skew
③	Uniformity
④	Subfigure uniformity
⑤	Maximum uniformity difference
⑥	Entropy ratio
⑦	Residual pixels
⑧	Skew difference
⑨	Variance difference
①⓪	Distance between centres of gravity

multi-model figure classifier achieved the smallest variance over all figure categories.

4. Discussion

4.1. Image features

Figures in bioscience literature are complex. In this paper, we explored both common image features and new image features we derived from biomedical subfigures, and experimented different image-feature combinations to identify the best set of image features for biomedical figure classification. Our results show that the eight new image features HIF derived from biomedical subfigures improve figure classification.

As shown in Table 1, HIF outperformed common image features (i.e., intensity histogram (IH) and edge directional histogram (EDH)). In addition, the results show that HIF performed consistently in different machine-learning classifiers, while other features varied significantly. For example, EDH showed inconsistent performances for four machine-learning classifiers, obtaining the highest accuracy with Naïve Bayesian (NB) model while performing the worst with artificial neural network (ANN). We also observed inconsistent performance when EDH was added as additional features. We speculate that such inconsistency may be caused by too small a training size. We therefore excluded EDH leading to the optimized feature combination IH + HIF. The results demonstrated that HIFs are robust image features for biomedical figure classification.

Most of previous work use moments and entropy for biomedical figure classification [14,15,18,19]. In contrast, our work (Fig. 3) has shown that those three image features do not strongly associated with any specific figure types. As a result, it is not surprising, as shown in Table 1, that when those three features are removed, the resulting IH* + HIF achieved improved performance by 3.9%.

We have identified the best image feature combination IH* + HIF, which consists of 10 distinctive image features as shown in Table 2. Based on these 10 image features, our image-feature based classifier (SVM (RBF) in Table 3) achieved an accuracy of 65% which is 11% (absolute value) higher than our previous work [15] and 30.6% better than the baseline system, as shown in Table 6.

4.2. Image features vs. text features

We explored three feature types – image features, text features, and the joint features (image + text features). Table 3 shows that SVMs performed the best for all feature types, although different kernel functions performed differently with different feature types. Since different feature types have built up different feature space,

Table 3
Performances of the machine-learning classifiers according to the feature types.

	Image feature (%)	Text feature (%)	Joint feature (%)
Machine learning-model	Accuracy (F1-score)	Accuracy (F1-score)	Accuracy (F1-score)
ANN	53.0 (48.3)	64.8 (56.7)	71.4 (71.3)
KNN	61.6 (47.0)	54.3 (40.6)	58.5 (49.9)
NB	58.0 (47.5)	15.4 (5.3)	15.4 (5.3)
SVM (linear)	43.6 (39.4)	**71.0 (69.6)**	72.3 (71.4)
SVM (polynomial)	48.3 (41.2)	64.0 (59.6)	**74.4 (72.8)**
SVM (RBF)	**65.0 (60.0)**	36.0 (13.5)	34.5 (10.3)
SVM (sigmoid)	34.5 (10.3)	36.0 (13.5)	34.5 (10.3)

Image feature: 10 image features in Table 2.
Text feature: 568 text features described in Section 2.2.3.
Joint feature: the combination of image and text features.
Bold values represent the best among figure classifiers for each feature type.

Table 4
Classification performance on different figure categories (best classifier for each feature type).

	Image feature (%) (SVM (RBF))	Text feature (%) (SVM (linear))	Joint feature (%) (SVM (polynomial))
Figure type	F1-score (recall, precision)	F1-score (recall, precision)	F1-score (recall, precision)
Gel-image	73.4 (79.7, 68.1)	65.6 (67.8, 63.5)	71.0 (74.6, 67.7)
Image-of-thing	82.5 (88.9, 76.9)	78.5 (68.9, 91.2)	82.9 (75.6, 91.9)
Graph	75.2 (89.4, 64.8)	78.0 (84.9, 72.3)	82.1 (87.1, 77.7)
Model	16.3 (9.2, 70.0)	78.4 (76.3, 80.6)	75.3 (80.3, 70.9)
Mix	52.5 (52.1, 52.9)	47.7 (43.7, 52.5)	52.5 (43.7, 66.0)

Table 5
Figure classification performances according to the figure classification models.

	Hierarchical (%)	SVM (%)	Multi-model (%)
Figure type	F1-score (recall, precision)	F1-score (recall, precision)	F1-score (recall, precision)
Gel-image	75.6 (81.4, 70.6)	71.0 (74.6, 67.7)	75.6 (81.4, 70.6)
Image-of-thing	86.0 (95.6, 78.2)	82.9 (75.6, 91.9)	86.0 (95.6, 78.2)
Graph	65.1 (71.2, 59.9)	82.1 (87.1, 77.7)	84.4 (86.4, 82.6)
Model	27.5 (23.7, 32.7)	75.3 (80.3, 70.9)	73.1 (69.7, 76.8)
Mix	62.2 (52.1, 77.1)	52.5 (43.7, 66.0)	64.5 (56.3, 75.5)

Table 6
Comparison of the overall figure classification performances.

Classification model	Accuracy (%)	F1-score (%)
Baseline system	34.4	51.2
Hierarchical classifier	62.7	63.3
SVM classifier	74.4	72.8
Multi-model classifier	77.8	76.7

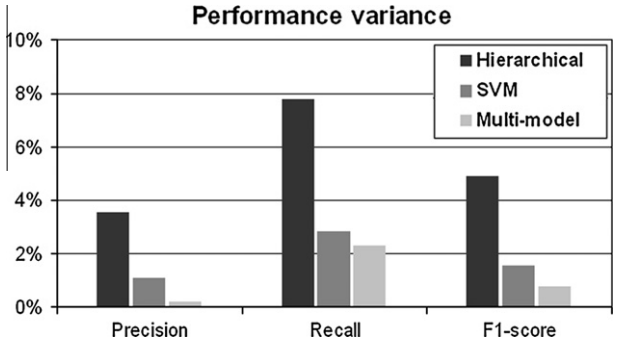

Fig. 10. Comparison of classification performance variation. Multi-model figure classifier presents the smallest variation among figure classifiers, namely, it is the most stable figure classification system.

the variation of feature space may result in varied performance by different kernel functions.

Consist with the previous studies [14,15], text-feature based classifier outperformed the image-feature based classifier 6% (absolute) in accuracy, as shown in Table 3. Interestingly, the results also show that the top 10 image feature-based classifier outperformed text feature-based classifier for three figure types: *Gel-image*, *Image-of-thing*, and *Mix*. The results suggest that there may be a room to further improve figure classification and we will investigate this in our future work.

Consistent with the previous work [15], our results show that the best figure classifiers were those that integrated both text and image features. As shown in Table 3, the joint-feature based classifier achieved an accuracy of 74.4% which is 3.4% and 9.4% higher than the text-based and image-based classifiers, respectively. In particular, it is 40% higher than the baseline system (the highest frequently occurred type (*Graph*) explained in Section 3.4) as shown in Table 6.

4.3. Hierarchical figure classifier vs. machine-learning classifier

We explored three classification models – hierarchical figure classifier, machine-learning classifier, and multi-model classifier. Table 6 shows that the SVM classifier outperformed the hierarchical classifier by 12.3% (accuracy) and 9.5% (F1-score), respectively. However, the SVM classifier did not outperformed the hierarchical classifier in all five figure types. As shown in Table 5, the hierarchical classifier outperformed SVM for *Gel-image* and *Image-of-thing*. We speculate that the reason is that many image features (six of 10 image features in Table 2) we identified are strongly associated with *Gel-image* and *Image-of-thing*, as mentioned in Section 2.3.1. The results also suggest the contribution of our new image features derived from subfigures.

Our results show that the multi-model classifier is the best system as it integrated the hierarchical and SVM classification models. As shown in Table 5, the multi-model classifier performed as well as the hierarchical classifier for *Gel-image* and *Image-of-thing* and outperformed SVM for *Model* and *Mix* to attain an overall of 77.8% accuracy and 76.7% F1-score, which are 15.1% (absolute value in accuracy) and 13.4% (absolute value in F1-score) and 3.4% (absolute value in accuracy) and 3.9% (absolute value in F1-score) higher than the hierarchical and SVM classifiers, respectively. As shown in Table 6, the multi-model classifier outperformed the baseline system 43.4% (absolute value in accuracy) and 25.5% (absolute value in F1-score).

4.4. Challenges and future work

Throughout our work, we identified four main challenges for biomedical figure classification. First, image feature selection was a challenge task. In this study, we employed heuristic approaches for image feature selection. Although the 10 image features we selected yielded improved performance, we speculate that additional image features can be explored and this remains our future work.

Secondly, the rule-based figure classifier was also a challenge task. All the presented rules and threshold values were heuristic, and as a result, it was very difficult to find the optimal rules and threshold values. Therefore, we need to improve the flexibility of

the rules through the use of other kind of rules (e.g., fuzzy rules) in our future work. In addition, we may automatically adjust the parameters of the rules to eliminate the subjectively associated to the heuristic design of the rules.

Thirdly, separating *Graph* from *Model* is challenging as the two figure types exhibited similar image feature patterns. Although text features in the figure captions were more distinctive than image features for *Graph* and *Model*, we should extract more image features of both *Graph* and *Model* for further improvement in the future.

The last challenge is the identification of *Mix*. Currently, we first segmented subfigures and extracted subfigure image features and then compared the image feature differences among subfigures and explored the differences for figure type classification. Although the performance for *Mix* has been improved in our study, more work is needed. One such future research direction is subfigure clustering [17].

5. Conclusion

In this study, we presented a figure classifier that automatically classifies biomedical figures into five predefined figure types. We explored rich image features and performed feature selection and the results show that the new image features and feature selection improved figure classification. The best figure classifier integrated both text and image features, and integrated both rule-based and supervised machine-learning classifiers, yielding an overall performance of 77.8%, which was 40% higher than the baseline system and 3.8% better than the state-of-the-art figure classifier [15]. Future work will focus on identifying *Mix* and separating *Graph* from *Model*.

Acknowledgments

We acknowledge the support of 1R01GM095476 to Hong Yu. Any opinions, findings, or recommendations are those of the authors and do not necessarily reflect the views of the NIH. We thank Lamont Antieau and Feifan Liu for editing the article. We also thank Lisha Choubey for the inter-annotation of figure data.

References

[1] Yu H, Lee M. Accessing bioscience images from abstract sentences. Bioinformatics 2006;22:547–56.
[2] Yu H, Agarwal S, Johnston M, Cohen A. Are figure legends sufficient? Evaluating the contribution of associated text to biomedical figure comprehension. J Biomed Discov Collaboration 2009;4.
[3] Agarwal S, Yu H. FigSum: automatically generating structured text summaries for figures in biomedical literature, AMIA annual symposium; 2009.
[4] Yu H, Liu F, Ramesh BP. Automatic figure ranking and user interfacing for intelligent figure search. PLoS One 2010;5.
[5] Hearst MA, Divoli A, Guturu H, Ksikes A, Nakov P, Wooldridge MA, et al. BioText search engine: beyond abstract search. Bioinformatics 2007;23:2196–7.
[6] Kahn CE, Thao C. GoldMiner: a radiology image search engine. Am J Roentgenol 2007;188:1475–8.
[7] http://www.biomed-search.com/.
[8] Xu S, McCusker J, Krauthammer M. Yale Image Finder (YIF): a new search engine for retrieving biomedical images. Bioinformatics 2008;24:1968–70.
[9] Murphy RF, Velliste M, Yao J, Porreca G. Searching online journals for fluorescence microscope images depicting protein subcellular location patterns. In: IEEE international symposium on bioinformatics and biomedical engineering; 2001. p. 119–28.
[10] Qian Y, Murphy RF. Improved recognition of figures containing fluorescence microscope images in online journal articles using graphical models. Bioinformatics 2008;23:569–76.
[11] Ahmed A, Xing E, Cohen W, Murphy RF. Structured correspondence topic models for mining captioned figures in biological literature. In: International conference on knowledge discovery and data mining; 2009. p. 39–47.
[12] Murphy RF, Kou Z, Hua J, Joffe M, Cohen W. Extracting and structuring subcellular location information from on-line journal articles: the subcellular location image finder. In: International conference on knowledge sharing and collaborative engineering; 2004. p. 109–14.
[13] Ahmed A, Arnold A, Coelho LP, Kangas J, Sheikh AS, Xing E, et al. Structured literature image finder. In: Proceedinfgs of the ISMB BioLINK special interest group; 2009. p. 209–12.
[14] Shatkay H, Chen N, Blostein D. Integrating image data into biomedical text categorization. Bioinformatics 2006;22:446–53.
[15] Rafkind B, Lee M, Chang S, Yu H. Exploring text and image features to classify images in bioscience literature. In: Proceedings of the BioNLP workshop on linking natural language processing and biology; 2006. p. 73–80.
[16] Cohen J. A coefficient of agreement for nominal scales. Educ Psychol Measur 1960;20:37–46.
[17] Gonzalez R, Woods R. Digital image processing. 2nd ed. Upper Saddle River, New Jersey: Prentice Hall; 2002.
[18] Tian YH, Huang TJ, Wen G. Exploiting multi-context analysis in semantic image classification. J Zhejiang Univ Sci 2005;6:1268–83.
[19] Kim D, Yu H. Hierarchical image classification in the bioscience literature. American Medical Informatics Association (AMIA) Symposium; 2009.
[20] Tamura H, Mori S, Yamawaki Y. Texture feature corresponding to visual perception. IEEE trans on systems, Man, and cybernetics Smc-8; 1978. p. 460–73.
[21] Jain K, Farroknia F. Unsupervised texture segmentation using gabor filters. Pattern Recogn 1991;24:1167–86.
[22] Jain K, Vailaya A. Shape-based retrieval: a case study with trademark image data bases. Pattern Recogn 1998;31:1369–90.
[23] Kou Z, Cohen WW, Murphy RF. Extracting information from text and images for location proteomics. ACM SIGKDD workshop on data mining in bioinformatics; 2003.
[24] Sable C, Hatzivassiloglou V. Text-based approaches for non-tropical image categorization. Intl J Digit Libr 2000;3:261–75.
[25] Onix Text Retrieval Toolkit. <http://www.lextek.com/manuals/onix/index.html>.
[26] Jones KS, Willet P. Readings in information retrieval. San Francisco: Morgan Kaufmann; 1997.
[27] Manning CD, Raghavan P, Schütze H. Introduction to information retrieval. Cambridge (UP); 2008. p. 253–87 [chapter 13].
[28] Majumder P, Mitra M, Chaudhuri BB. N-gram: a language independent approach to IR and NLP. In: International conference on universal knowledge and language; 2002.
[29] Sato S, Hayashi H, Maki N, Inoguchi M. Development of automatic keyword extraction system from digitally accumulated newspaper articles on disasters. In: International conference on urban disaster reduction; 2007.
[30] Liu F, Liu Y. Automatic keyword extraction for the meeting corpus using supervised approach and bigram expansion. IEEE spoken language technology; 2008.
[31] Cascia ML, Sethi S, Sclaroff S. Combining textual and visual cues for content-based image retrieval on the world wide web. IEEE workshop on content-based access of image and video libraries; 1998. p. 24–8.
[32] Schölkopf B, Platt J, Shawe-Taylor J, Smola AJ, Williamson RC. Estimating the support of a high-dimensional distribution. Neural Comput 2001;13:1443–71.
[33] http://opencv.willowgarage.com/wiki/.