A Semi-Automatic Indexing System for Cell Images

Chaoxin Zheng, Dermot Kelleher and Khurshid Ahmad

Abstract—A method is described that can be used for annotating and indexing an arbitrary set of images with texts collateral to the images. The collateral texts comprise digitised texts, e.g. journal papers and newspapers in which the images appear, and digitised speech, e.g. a commentary on the contents of the images. The annotation 'vector' comprises image features and keywords in the collateral texts; our method can be used to generate both the image features and keywords automatically. Terminology extraction techniques are incorporated into the system to form a domain specific lexicon, which can then be used or help to annotate the images. Our method can be used as the basis of the autonomous learning of associations between images and their collateral descriptions, for example using Kohonen feature maps. We focus on images that show the migration and the division of cells within live systems. We show how the annotations can be collected by using the state-of-the-art speech recognition techniques that convert audio input into descriptive text on cell migration. A system based on the method has been developed and has reduced the annotation time to around two minutes per image, on a set of 429 cell images − which is significantly smaller than 5 minutes for manual annotation.

I. Introduction

ADVANCES in microscopy and related image analysis techniques is enabling molecular biologists and others to visualise molecular and morphological events inside cells and tissues [1]. High content screening (HCS) systems [2], comprising high-magnification fluorescent microscopy systems and sophisticated image processing/analysis techniques [3], facilitate the 'measurement of changes in objects locations, intensity, texture or shape, thereby enabling the analysis of more subtle and physiological relevant cellular events such as cell or protein movements, shape changes or protein modification' [4]. HCS systems are particularly important for the study of 'responses to chemical or genetic perturbations within the context of the inter-cellular structure and functional networks of individual normal or diseased cells' [1]. HCS systems are critical to the needs of both the academic and pharmaceutical research in that these systems facilitate the measurement of cytological changes, the selection of cell populations, and facilitate the combination of the measures related to the various components of a cell, for example the cell nucleus structure and the chromatin structure [3].

Whilst the interpretation of images is still the preserve of human experts, the massive throughput of the HCS systems cannot be managed by manual archiving of the images, for instance, in the file directories available on the personal and laboratory computers. For one, the images have to be indexed

Chaoxin Zheng and Khurshid Ahmad are with the Computer Science, Trinity College Dublin, Dublin2, Ireland (email: {chaoxin.zheng, khurshid.ahmad}@cs.tcd.ie)

Dermot Kelleher is with the Clinical Medicine, Trinity and Dublin Molecular Centre, St. James. Hospital, Dublin8, Ireland.

for subsequent retrieval and for this purpose one has to rely on ready access to an up to date terminology, and inevitably that of the experts. There is clearly a need for systems that can learn to associate images with keywords automatically [5]–[8] and a prototype system for the automatic annotation of cell images has recently been reported. We have learnt that information in visual features can be complementary to that contained in the keywords; for example, a cell belonging to a specific patient undergoing a specific treatment can be described best by its visual features and the properties of the cell, its genus, its part-whole and causal relationship can be described best by using linguistic features like keywords [9]. The semantic 'gap' [10], between an image's visual feature and its linguistic description, is not a 'gap' for us, rather two features that are essential for the description of an image as a whole.

In this paper we report on the developments of our CellLab system which comprises a cropus manager for storing and retrieving (cell) images, a component for extracting key visual features, especially segmentation of cell images using hierarchical SOM's, and another component for terminology extraction and annotation. Finally the CellLab has a learning system that learns to associate images features with linguistic descriptions of an image using a multinet neural network system.

Our focus will be on terminology extraction from collateral text that does not rely on extant terminology systems but does learn the key terms through an automatic/contrastive analysis of learned journals. Furthermore, we have been experimenting on using speech-to-text technologies, that can recognise speech fragments like keywords, for image annotation: there are two reasons for this. First, the number of images that are published in scientific journals is limited due to cost and space, although both limitations have been alleviated due to electronic publishing. Second, the future scenario we envisage is as follows: busy molecular biologists will be using HCS systems and using a lapel microphone will describe the contents of the important images. The audio description of the images will form another useful corpus of collateral texts in addition to journals and books.

II. Motivation and Literature Review

A. Motivation

Automatic image indexing is still an open challenge in computer vision and in artificial intelligence, and it is the key to the successful development of multimedia retrieval systems. Many intelligent system have been built to learn to automatically establish the link between images and linguistic. One important step in this enterprise is the collection and creation

of a huge exemplar set which comprises images and their collateral text. Much of the effort here is on large collections of images of everyday objects behaving spontaneously in natural or built environments − the erstwhile Corel dataset is a good example of such undifferentiated datasets which has been used extensively in benchmarking of image indexing systems. If we look at our own domain of study, cell migration, there are indeed some sophisticated databases available online, but their annotation leaves much to be desired. The objective of the paper is to describe the set up of a semi-automatic annotation system, which enable some experts in a specific domain to efficiently annotate images that fall in their expertise.

B. Literature Review

Volkmer reports on [11] a web-based system called EVA was developed for collecting manual annotation of large image and video collections by IBM Corp for research. EVA was designed based on for promoting 'consistency, simplicity, and the speed of annotation'. However, EVA developers have made three key assumptions assumptions: all annotations are selected from a small and controlled vocabulary; annotations are for static visual concepts (images) only; and annotations are generated for the entire video clip (a set of similar images) at the global frame level. During every annotation session, a typical EVA user is asked to select a few concepts provided by the system, which they think are related to the video or image content. This will place a significant burden on the annotator and will make the process quite subjective.

Aria is another system for annotating and retrieval of images developed by the MIT Media Labs [12]. Unlike the EVA, Aria is an agent, which uses a quite different way of generating annotation for images. Instead of browsing images and attaching keywords to the images as EVA does, Aria first presents the keywords and provides a browser to select images that are supposed to be related to the keywords presented. The annotations are then automatically attached to the image and stored in the database for retrieval later. The purpose of the development of Aria agent is that 'The users types the description of picture in order to communicate to another person, not specifically to image annotation. But once the computer has the input, why not take advantage of it'. The Aria system appears to place a lesser burden on its users, say, when compared to the EVA system.

Another way of building annotated image database is to make use of the internet search engine, such as Google. Images are collected from the internet, and the annotations for the images are the collateral text that surrounds the image. A given keyword is used to retrieve images and their associated textural description from a large-scale web image database [13]. Since keywords might be ambiguous, content-based matching is used to filter the visually irrelevant images and generate a ranking. Elsewhere, intelligent agents have been used to conduct hyper link-based search to collect images, and perform two types of classification to separate high-level and low-level semantic concepts from a pre-build ontology [14]. Such systems appear to generate an annotation database

without any human intervention, which means that the whole procedure can becomes fully automatic. However, it is not clear how proximate the terms are to an image and as such there is no guarantee that the images are related to the textural description. The amount of unrelated images and collateral text are quite huge, which will affect the learning procedure.

These annotation systems are based on natural images and are quite easy to understand and annotate, i.e. most people are able to read the image and attach keywords to them. In this paper, we are dealing with cell images, which are generated by high throughput microscopes, either fluorescence or transmitted light. The context of the images is indeed quite different from the natural images. We have specialist cell migration images and perhaps only experts in the area are eligible to analyse and annotate the images. The available resource where the image annotations came from are significant limited − only experts in the domain, and therefore it calls for the need of an efficient annotation system. In particular, a system which can make good use of the limited resource and produce high quality of annotations for representative cell images. Such images and their collateral annotation can then be learnt to annotate other 'unseen' images for image indexing and retrieval.

This paper is organised as follows: In the next section, we will outline the design and architecture of our annotation system, followed by the detail implementation including the audio to text conversion for annotation input. Based on the available annotation set, keywords related to the cell migration domain are generated by using linguistic processing. Such keywords, including compound terms if necessary, can then be used to fast annotate or index the images. Some key results from the annotation procedure are presented and discussed, and finally the conclusion and some future remarks on image annotation.

III. METHODS

A. Image Annotation

Our proposed system for generating a linguistic feature vectors from an arbitrary collection of (image collateral) texts can be integrated with an image feature analysis system. The inputs to, and outputs from, our linguistic processing system are shown in Fig. 1. The details of our image feature analysis and image vector generation has been presented in detail elsewhere [9]. The linguistic feature analysis and annotation system discussed in this paper has six modules that are essential for a practical system that can be used in a laboratory condition.

1) First: The annotation system has access to an image management system that can help an end user to manage annotation of new images and revision of annotations to existing images; the management system provides for routine housekeeping facilities like dates and times on which an image was first and subsequently annotated.

2) Second: A browsing facility has been provided for a quick-browse of (parts of) an image collection. Zooming and panning functions of our system help a user to view an image at high-level of resolution (as captured, say, from a high-level
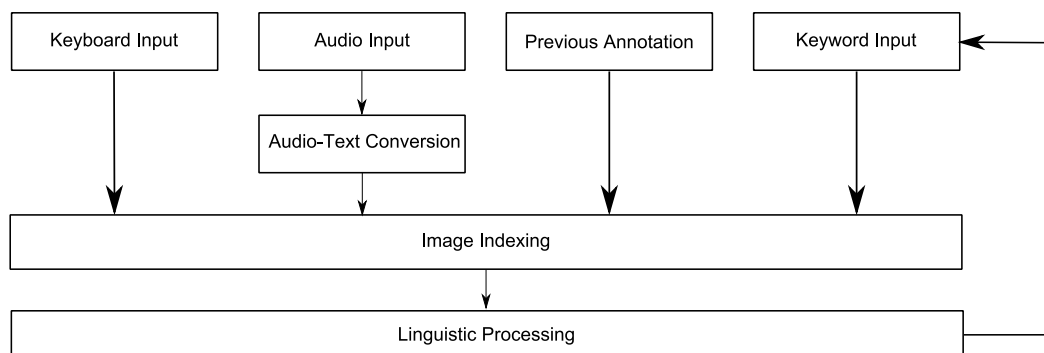
Fig. 1.   Architecture of the image annotation system.

resolution microscope) or at a coarser level of granularity to observe global features of an image collection.

3) Third: The linguistic feature vector can be (i) automatically generated from a set of images that appear embedded in a text document, and the user has a choice either to use the whole text or to select texts parts of are very closely collateral (for example, the captions of an image) or broadly collateral that is the names of the authors of the paper in which the image appears or the names of the authors of texts that have been cited in the paper, or (ii) generated from an audio description provided of the image by an expert in uncontrolled natural language. Our annotation system has interface with a speech-to-text convertor, specifically Dragon Naturally Speaking from Nuance Corp. (more about the analysis of free text/speech in the next section). It captures speech from the annotator through a noise reduction microphone and convert the speech into text instantly. The conversion is indeed quite simple so users can concentrate on the annotation rather than paying attention to the typing.

4) Fourth: The system allows for the user to add keywords himself or herself by allowing keyword input from keyboard for authorised users. Furthermore, the user may recall that the keywords to be used for a new image are essentially the same as that of an annotated image already in the system, for instance, cell mitosis images do comprise many frames and a number of frames before or after the mitosis may be exactly the same. There may be instances where a coarse image of a cell is presented in a paper or captured through a microscope followed by a finer resolution image: the coarser grained images may act as a genus − features in the coarse grained image may be shared by images at finer grain. This facility, we have observed, speeds up the annotation procedure quite considerable.

5) Fifth: There is an aide memoir feature to note (in free text) that there are features of an image which is being annotated that remind the user, for example, of another image, another technique, or another published paper. And

6) Sixth: There are some other features to enhance the annotation/indexing process. For instance, the images can be
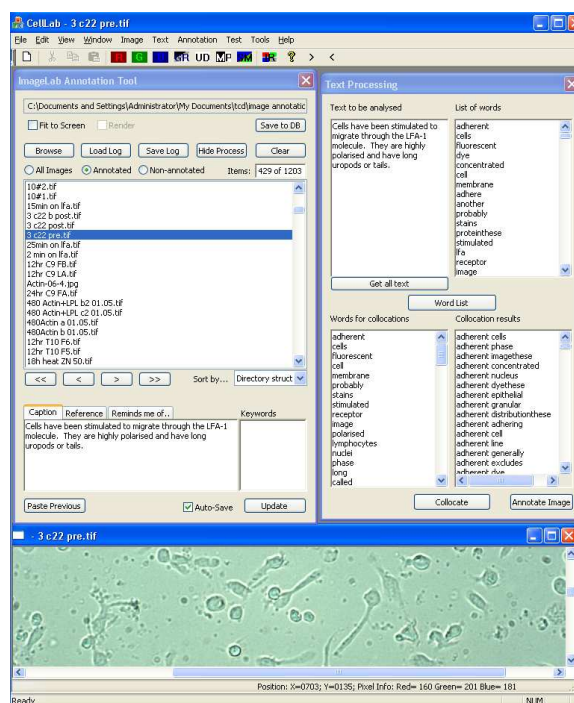

Fig. 2.   Illustration of the image annotation system CellLab.

sorted using different index including the storage structure (e.g. folder structure), image names, last time when the image is modified, and the size of the *caption* (annotation). A snapshot of our developed system CellLab is illustrated in Fig. 2.

B. Automatic Linguistic Processing

A lexicon, a collection of words, is critical for a number of natural language processing system (NLPS) including computational linguistics, information retrieval systems, infor-

mation extraction systems and machine translation systems. In computational linguistics Image annotation is one of the newer applications of NLPS. Much of the effort has been It is well known that there is a lexical 'bottleneck' in information extraction and in computational linguistics [15]. Most natural language processing literature is dedicated to language of everyday usage which is well known for its imprecision and uncertainty: there is ambiguity at the grammatical level as well. The specialist language texts are rather more constrained and rely on a relatively larger vocabulary but a restricted grammar and the vocabulary usage shows characteristic patterns for different specialist domains. These characteristics can be identified by comparing the distribution of single word vocabulary in a collection of specialist language texts with that of the same words in a systematically organised collection of general language texts [16]. The application of this method of contrasting two collections of texts has been outlined for knowledge acquisition has been discussed in the literature [17]. It is the specialist texts that interests us, specifically the domain of cell migration.

One key characteristic of specialist texts is the extensive use of compound or multi-word terms and these terms, in principle, can be selected by using general principles of word formation, including the interaction of morphology and syntax [18], [19]. One complementary technique of identifying and extracting compound terms is the use of techniques in statistical linguistics [20], [21] and compute not only the frequency of distribution of single words but also that of words occurring together. Such an approach has been used to successfully extract multi-word terms. The multi-word terms can be organised into a hierarchy, based on the assumption that in English the headword is assumed to be the rightmost word in a compound − the hierarchy is part of the ontology of the domain. So the use of a contrastive analysis of the distribution of words, and the co-occurrence of the most salient single words, can be used to build a specialist lexicon and be use to extract the ontology of the domain itself [22].

The above mentioned methods for extracting single words and compound words have been implemented for extracting terms from an arbitrary collection of texts for image annotation purposes.

1) Frequency Analysis: For the extraction of single keyword, frequency distribution of the words and statistics associated with the frequency are analysed. One prerequisite step in extracting the terminology in cell migration is to remove all the grammatical words such as determiners and repositions, which are irrelevant to the special domain. This is conducted by creating a stop list. Two measurements are then calculated for every word in the corpus, frequency f (or relative frequency: the frequency of the word over the total number of words in the corpus), and the weirdness w, which is computed using:

$$w = \frac{f_s/N_s}{f_g/N_g} \tag{1}$$

where s stands for the special corpus, cell migration in our case, g means the general corpus − we use the British

Language Corpus, and N is for the total number of words in the corpus. The relative frequency shows how many times that a word occurs in the corpus, and the more it occurs, the more important it is. A high value of the weirdness represents that the word is much more frequently used in the special language domain than it is in the general corpus, and thus the word is much more important for the special language domain. The frequency and weirdness of all the words is normalised using the so-called z-score to be with the average of 0 and the standard deviation of 1. A word is then selected as the keyword, or the lexicon unit, if and only if the z-scores of both frequency and weirdness are above the predefined thresholds, which are both 0 here.

2) Collocations: Collocations are used to extract compounds. For every keyword a window of five words either side the keywords is examined as a potential two, three, four or five word compound. The significance of the collocations is measured in terms of mutual information related statistics, including the strength, spread, and frequency. The selection criteria are described as [20]:

$$strength = \frac{f - \overline{f}}{\sigma} \geq t_0 \tag{2}$$

$$spread \geq U_0 \tag{3}$$

$$f_{c_j}^i \geq f_{c_j} + (k_1 \times \sqrt{U_i}) \tag{4}$$

where k_0, k_1, and U_0 are three pre-selected thresholds to filter the the statistical insignificant collocations, and following Smadja are set to be 1, 1, and 10, respectively here. Parameter f is the frequency of collocation between every two words, and f_c means the frequency of each collocation at different positions usually from five to the left and five to the right. As explained in [20], the equation 2 is used to 'filter out irrelevant data, that is pairs of words supposedly not used consistently with a sigle syntactic structure', equation 3 'requires that the histogram of the 10 relative frequency ... to have at least one spike', and equation 4 'pulls out the interesting relative positions of the two words'.

IV. RESULTS AND DISCUSSION

A. Image Annotation

Three experts in cell migration have tested our system. They browsed one image at a time; and used the speech input facility was used principally to describe the image The three experts spent 13 hours , over a period of 13 days, annotating the images: it took an additional one hour per expert for training the speech recognition software - these software systems are dependent on the speakers voice to an extent. A total of 429 images were annotated. Four example images with their annotations are presented in Fig 3.

We have noted that during the annotation, quite a large amount of time was spent on handling and training the speech recognition, particularly at the beginning, as the performance of the software appeared to be very low at the beginning, − there were quite a lot of errors in the detection, so the experts had to spend much time in correcting the detection. This was
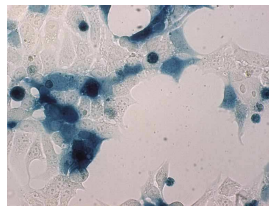
(a) These are adherent cells labeled with a blue dye, probable trypan blue. This dye is excluded from viable cells but is taken up by dead or dying cells. This is a bright field image.
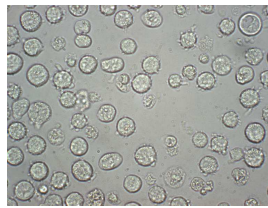
(b) These cells were acquired using transmitted light they seem to be in poor health exhibiting membrane blebbing small membrane fragments can be seen lying adjacent cells.
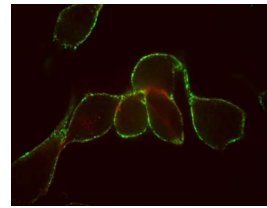
(c) Human epithelial cells stained to detect surface distribution of bacterial receptor. Localisation of the receptor is demonstrated by Green fluorescent staining around the cell membrane.
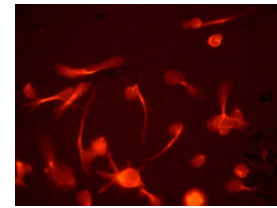
(d) Human lymphocytes activated by antibodies to cell surface receptor LFA-1. Red staining shows micro-tubules.

Fig. 3.   Four examples of cell images and their annotations.

TABLE I
FREQUENCY ANALYSIS OF ANNOTATION OF CELL IMAGES

Word	Frequency	Relative F	Weirdness
dye	194.000	0.017	3905.429
fluorescent	127.000	0.011	3711.809
nuclei	110.000	0.010	3162.421
nanoparticles	52.000	0.005	457458.578
dapi	46.000	0.004	40467.490
granular	44.000	0.004	5692.358
adherent	40.000	0.004	4755.287
microtubules	35.000	0.003	23684.986
trypan	32.000	0.003	21654.844
polygonal	31.000	0.003	5681.577
phagocytic	29.000	0.003	31890.141
actin	29.000	0.003	3543.349
reticulum	24.000	0.002	4910.110
localisation	24.000	0.002	3248.227
phagocytes	19.000	0.002	8797.280
endoplasmic	19.000	0.002	5391.881

perhaps due to that there was a large number of technical terms and abbreviations used in the special domain. However, as more images are annotated the system learns more terms and the annotation process as a whole is speeded up.

B. Terminology Extraction

We have analysed the annotations collected for a total of 429 images, and results are demonstrated in Table I. The corpus comprised a total of 11379 words and a vocabulary of 574 words −; a larger more representative corpus is being constructed currently. It should be noted from the table that although the keyword *dye* and *fluorescent* are the two most frequent ones, their weirdnesses are not the highest. Instead, *nanoparticles* have the highest weirdness of all the words including the words unlisted, and it therefore is one of the most significant keywords for the domain of cell migration. Another quite interesting finding here is that the word *cell* is excluded from the lexicon and not listed here. Although the word *cell* including *cells* has been used for 735 times, which is the most frequent word in the corpus, but its z-score of the weirdness (-0.09) is below the threshold (0) and thus rejected from the lexicon. Other words such as *phagocyte*, *colocalisation*, *uropod*, and *cytoskeleton* have indeed quite large weirdnesses. However, their occurrences in the corpus are not statistical significant for them to be included in the lexicon.

The single terms listed in Table I are further used for the collocations. After the first collocation two-word compounds are extracted and again used for the collocations and selection of three-word compounds. Some compounds organised in a tree structure are displayed by Portégé (Version 3.0, Stanford University) and presented in Fig. 4. In the collocations, some important compounds in the cell migration are included into the lexicon such as *actin faliment*, *actin cytoskeleton*, and *complex nuclei*. Recall that the term *nanoparticles* is the one with the largest weirdness, which indicates that it is quite an important term in the domain. However, as shown in Fig. 4, there are only two statistically significant two-word compounds detected in the corpus. This is possibly due to the fact that the corpus is still not large enough. Therefore, linguistic and technical experts are required as a complement to the automatic terminology extraction. As the size of corpus increases with more images annotated, the system will be able to provide terms with less and less interventions of the experts.

V. CONCLUSION AND FUTURE WORKS

In this paper, a semi-automatic image indexing system for cell migration study is described. A user-friendly interface is developed and employed to collect manual annotations from experts in the domain for a large set of cell migration images. The system allows free text input, converted from audio input from microphone, and thus provides the user with more freedom in selecting the proper annotation than many other annotation and indexing systems. Annotations are then processed and a lexicon, consisting of simple and complex terms, is extracted and presented back to the users to help during the annotation. However, it has been noted that the annotation corpus is still needs expansion, and we are currently in the progress of collecting more annotations. The automatic annotation and indexing of images based on the existing annotations remains our future works. Once we have collected enough annotations, our developed cross-modal system [9] will be applied to learn the associations between the image and the linguistic (annotation) features, which can
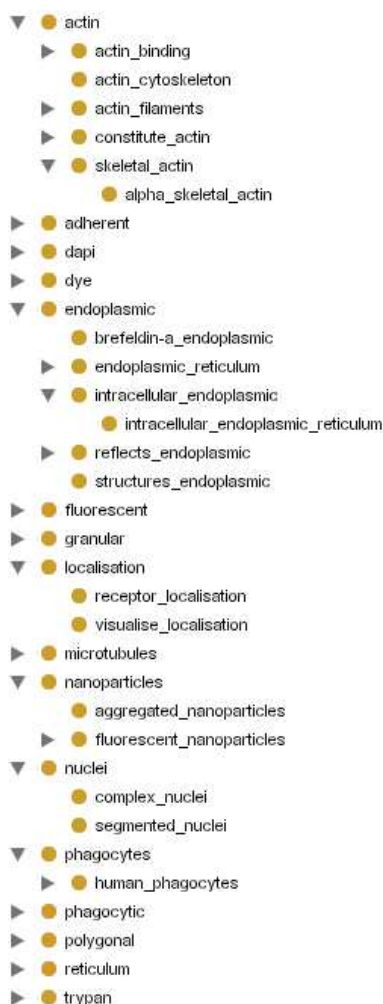
- ▼ 🟠 actin
  - ▶ 🟠 actin_binding
  - 🟠 actin_cytoskeleton
  - ▶ 🟠 actin_filaments
  - ▶ 🟠 constitute_actin
  - ▼ 🟠 skeletal_actin
    - 🟠 alpha_skeletal_actin
- ▶ 🟠 adherent
- ▶ 🟠 dapi
- ▶ 🟠 dye
- ▼ 🟠 endoplasmic
  - 🟠 brefeldin-a_endoplasmic
  - ▶ 🟠 endoplasmic_reticulum
  - ▼ 🟠 intracellular_endoplasmic
    - 🟠 intracellular_endoplasmic_reticulum
  - ▶ 🟠 reflects_endoplasmic
  - 🟠 structures_endoplasmic
- ▶ 🟠 fluorescent
- ▶ 🟠 granular
- ▼ 🟠 localisation
  - 🟠 receptor_localisation
  - 🟠 visualise_localisation
- ▶ 🟠 microtubules
- ▼ 🟠 nanoparticles
  - 🟠 aggregated_nanoparticles
  - ▶ 🟠 fluorescent_nanoparticles
- ▼ 🟠 nuclei
  - 🟠 complex_nuclei
  - 🟠 segmented_nuclei
- ▼ 🟠 phagocytes
  - ▶ 🟠 human_phagocytes
- ▶ 🟠 phagocytic
- ▶ 🟠 polygonal
- ▶ 🟠 reticulum
- ▶ 🟠 trypan

Fig. 4. An illustration of some compounds in cell migration domain.

be used to index 'unseen' images.

VI. ACKNOWLEDGMENT

REFERENCES

[1] K. Korn and E. Krausz, "Cell-based high-content screening of small-molecule libraries," *Current Opinion in Chemical Biology*, vol. 11, 2007.

[2] X. Zhou, X. Cao, Z. Perlman, and S. T. C. Wong, "A computerized cellular imaging system for high content analysis in monastrol suppressor screens," *Journal of Biomedical Informatics*, vol. 39, no. 2, pp. 115–125, 2006.

[3] S. A. Haney, P. LaPan, J. Pan, and J. Zhang, "High-content screening moves to the front of line," *Drug Discovery Today*, vol. 11, 2006.

[4] O. Rausch, "High content cellular screening," *Current Opinion in Chemical Biology*, vol. 10, 2006.

[5] K. Ahmad, M. Tariq, B. Vrusias, and C. Handy, "Corpus-based thesaurus construction for image retrieval in specialist domains." in *Proceedings of the 25th European Conference on Information Retrieval Research*, ser. Lecture Notes in Computer Science, F. Sebastiani, Ed., vol. 2633. Springer, 2003, pp. 502–510.

[6] L. Cheng and A. Sanchez-Azofeifa, "Component optimization for image understanding: A bayesian approach," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 28, no. 5, pp. 684–693, 2006, fellow-Terry Caelli.

[7] J. Li and J. Z. Wang, "Automatic linguistic indexing of pictures by a statistical modeling approach," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 25, no. 9, pp. 1075–1088, 2003.

[8] G. Carneiro, A. B. Chan, P. J. Moreno, and N. Vasconcelos, "Supervised learning of semantic classes for image annotation and retrieval," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 3, pp. 394–410, 2007.

[9] C. Zheng, A. Long, Y. Volkov, A. Davies, D. Kelleher, and K. Ahmad, "A cross-modal system for cell migration image annotation and retrieval," in *Proceedings of the 2007 International Joint Conference on Neural Networks*, 2007.

[10] A. W. M. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain, "Content-based image retrieval at the end of the early years," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 12, pp. 1349–1380, 2000.

[11] T. Volkmer, J. R. Smith, and A. P. Natsev, "A web-based system for collaborative annotation of large image and video collections: an evaluation and user study," in *MULTIMEDIA '05: Proceedings of the 13th Annual ACM International Conference on Multimedia*. New York, NY, USA: ACM, 2005, pp. 892–901.

[12] H. Lieberman, E. Rozenweig, and P. Singh, "Aria: An agent for annotating and retrieving images," *Computer*, vol. 34, no. 7, pp. 57–62, 2001.

[13] X.-J. Wang, L. Zhang, F. Jing, and W.-Y. Ma, "Annosearch: Image auto-annotation by search," in *CVPR '06: Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. Washington, DC, USA: IEEE Computer Society, 2006, pp. 1483–1490.

[14] L. Khan, "Standards for image annotation using semantic web," *Computer Standards & Interfaces*, vol. 29, no. 2, pp. 196–204, 2007.

[15] R. Basili, G. de Rossi, and M. T. Pazienza, "Inducing terminology for lexical acquisition," in *Proceedings of the Second Conference on Empirical Methods in Natural Language Processing*, C. Cardie and R. Weischedel, Eds. Somerset, New Jersey: Association for Computational Linguistics, 1997, pp. 125–133. [Online]. Available: citeseer.ist.psu.edu/550692.html

[16] K. Ahmad and M. A. Rogers, "Corpus linguistic and terminology extraction," in *Handbook of Terminology Management*, S.-E. Wright and G. Budin, Eds. Amsterdam & Philadelphia: John Benjamins Publishing Company, 2001, vol. 2.

[17] K. Ahmad, "The role of specialist terminology in artificial intelligence and knowledge acquisition," in *Handbook of Terminology Management*, S.-E. Wright and G. Budin, Eds. Amsterdam & Philadelphia: John Benjamins Publishing Company, 2001, vol. 2.

[18] B. Daille, Éric Gaussier, and J.-M. Langé, "Towards automatic extraction of monolingual and bilingual terminology," in *Proceedings of the 15th Conference on Computational linguistics*. Morristown, NJ, USA: Association for Computational Linguistics, 1994, pp. 515–521.

[19] B. Daille, "Study and implementation of combined techniques for automatic extraction of terminology," in *The Balancing Act: Combining Symbolic and Statistical Approaches to Language*, J. Klavans and P. Resnik, Eds. Cambridge, Massachusetts: The MIT Press, 1996, pp. 49–66. [Online]. Available: citeseer.ist.psu.edu/daille94study.html

[20] F. Smadja, "Retrieving collocations from text: Xtract," *Computational Linguistics*, vol. 19, no. 1, pp. 143–177, 1993.

[21] C. D. Manning and H. Schtze, *Foundations of Statistical Natural Language Processing*. Cambridge, Massachusetts: The MIT Press, June 1999.

[22] K. Ahmad, "Artificial ontologies and real thoughts: Populating the semantic web?" in *Proceedings of the 10th Congress of the Italian Association for Artificial Intelligence*, 2007, pp. 3–23.