Pattern Recognition Letters 72 (2016) 15–24

Contents lists available at ScienceDirect

Pattern Recognition Letters

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

Visipedia circa 2015☆

Serge Belongie[a], Pietro Perona[b,*]

[a] Department of Computer Science, Cornell University and Cornell Tech, United States of America
[b] Department of Electrical Engineering, Department of Computation and Mathematical Sciences, and Computation and Neural Systems Option, California Institute of Technology, United States of America

ARTICLE INFO

Article history:
Available online 10 December 2015

Keywords:
Visipedia
Visual recognition
Human-machine interaction
Machine learning
Active learning
Wikipedia
Visual psychology
Crowdsourcing
Computer Vision

ABSTRACT

Visipedia is a network of people and machines designed to harvest and organize visual information and make it accessible to anyone who has a visual query. We discuss technical challenges arising from Visipedia and discuss their implications for pattern recognition, computer vision, machine learning and visual psychology. Amongst these are discovering visual information that is implicit in experts' brains and in crowds of people and estimating its accuracy. To motivate our thinking we explore a case study, an automated field guide to the birds of North America. We conclude by discussing research directions that are necessary to make progress on Visipedia. An important realisation is that the study of 'computer vision' and 'machine learning' has to be broadened to include the process of information discovery and the dynamic interaction of people and machines in this context. Human-machine systems with no oracle are now within the scope of pattern recognition, machine learning and computer vision.

1. Introduction

In September 2014 Randall Munroe, the author of the popular web comic *xkcd*, published the panel shown in Fig. 1 highlighting "the difference between the easy and the virtually impossible" in Computer Science. The challenge proved irresistible for the computer vision and machine learning team at Flickr who, just one month later, released the "Park or Bird" web app demonstrating that the 'virtually impossible' was, in fact, possible (Fig. 2). Was Munroe proved wrong? Not quite: as he suspected, solving the park question was easy (use the geotag), while solving the bird part was very challenging. The Flickr team's success was enabled by a very recent computer vision and machine learning technical breakthrough [1,2].

As we play with Flickr's clever app, our curiosity is piqued: OK, great, it is a bird. But what *kind* of bird is it? "Bird," like "truck" or "flower," is an *entry level category* [3], at the level of abstraction most people commonly use to think and talk about an object. The species of the bird is called 'subordinate category' by psychologists and 'fine-grained category' by computer vision scientists. We all see the 'bird', but most of us can not recognize its species (we wish we could).

This sentiment, together with mild frustration with current technology, is echoed by Yahoo! president Marissa Mayer (interview in *IEEE Spectrum* in March of 2012):

> You see a bird and want to know what it is. You can take a picture of it, and Google's Goggles will tell you that it's a bird (but you already knew that, didn't you?).

Yes, indeed. Would it not be nice to have an app on our smart device, so that we can point our camera at any object and instantly learn about it in detail? Have we all not wished at some point that we could classify insects, architectural styles, animal bones and many other things? Mayer raises two important points. First, some of our queries are inherently visual. Wikipedia is wonderful, but what do we type into the search box when we wish to identify the bird that is pecking at our birdfeeder? Her second point is that, while there has been much progress in computer vision, the visual queries that are most interesting to humans are not yet addressed by machines.

This brings us to the goal of the Visipedia project, which is *to build a system for discovering and organizing visual information and making it easily accessible to anyone*. We argue that the right solution is similar to Wikipedia, albeit with a bigger role for automation – a decentralized, continually improving collaborative network of people and machines. Obviously, dealing with images is not as easy as dealing with text and thus Visipedia presents a number of unique and interesting challenges in pattern recognition, learning, psychology and beyond. In the remainder of this article we explore some of these challenges and describe our experience tackling a few

☆ This paper has been recommended for acceptance by Rama Chellappa.

* Corresponding author. Tel.: +1 (626) 395-4867.
E-mail addresses: sjb344@cornell.edu (S. Belongie), perona@caltech.edu (P. Perona).

http://dx.doi.org/10.1016/j.patrec.2015.11.023
0167-8655/

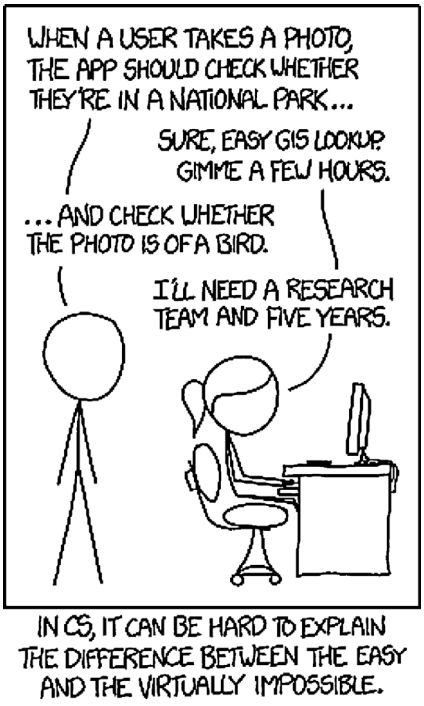

Fig. 1. Comic by Randall Munroe (http://xkcd.com/1425, September 2014).

of them. We will use bird identification as a case study throughout. We conclude with the observation that the scope of pattern recognition, computer vision and machine learning is broader than we thought.

2. A digital field guide for birds

2.1. Merlin, powered by Visipedia

In June 2015 we released an on-line field guide to North American birds called *Merlin*[1]. At the time of writing the system knows 400 out of about 1000 North American bird species and can classify them them from a picture. We engaged in this project in collaboration with the Cornell Laboratory of Ornithology to help us identify the issues that one will encounter in the more general Visipedia setting. Fig. 3 depicts a use case; the screenshots capture the user experience, which starts with photo submission and concludes with a brief description and an audio recording of the identified bird.

The computer vision architecture of *Merlin* consists of four steps [4] reflecting our current understanding of fine-grained classification [4–6]:

1. *Detection* – the bird is localized in the image using a 'bird' detector that is trained to work on all species.
2. *Part registration* – key landmarks (bill, belly, tail, feet, etc.) of a general-purpose 'bird model' are identified in the picture. The corresponding image patches are appropriately rectified.
3. *Feature extraction* – features are computed from the image using a reference frame that is defined by the landmarks.
4. *Fine-grained classification* – the features are used in a multi-class classifier to produce a shortlist of possible matches.

This architecture is quite sophisticated and contains a number of insights on how to represent features, how to compute pose for 3D objects, how to detect categories, etc.

While interesting, the computer vision bits are not the focus of Visipedia. From the point of view of Visipedia this is technology

[1] http://merlin.allaboutbirds.org/photo-id.

Fig. 2. Flickr's response to Munroe (Fig. 1). October 2014 (http://parkorbird.flickr.com).

which, we assume, will continue to be perfected over time by talented computer vision and machine learning researchers. The focus of the Visipedia project is understanding how humans and machines may best cooperate in discovering, organizing and searching visual information.

2.2. Humans and machines are complementary

A first realization we reached through developing and deploying *Merlin* is that the user and the machine have complementary abilities and mutually benefit from collaboration [7,8]. On the one side, the user is ignorant as to bird species and taxonomy (e.g., that the 'pileated woodpecker' is a species of woodpeckers, that it has a red crest, and is related to a number of South American woodpeckers). This is precisely why a user would engage with Visipedia. However, the user can see very well and will detect easily a 'bird' even when half hidden in a bush. He will also detect easily the main parts of the bird: eyes, tail, wings. Conversely, the machine has perfect knowledge of taxonomy and attributes, but cannot see as well as the user. Often (currently about 75% of the time [4]) the machine can detect and classify the bird in the picture. However, if it is confused the machine can ask the user for a hint, e.g., to click on the bill (see Fig. 3), which for the user is an easy and quick task. This simple information is valuable for the machine. It helps part registration (step (2) above), allowing it to read correctly the bird's attributes and complete the task successfully. Thus, *a machine collaborating with a human can solve a visual query that neither human nor machine could solve by themselves*.

Fig. 4 shows screen captures of a collaborative GUI for bird species classification [7,8]. Here a fast and accurate field guide was obtained by combining computer vision algorithms and human observers. Computer algorithms detect parts and predict attributes and human observers answer simple questions (e.g., 'what is the primary color of the bird') or perform simple actions (e.g., 'click on the head'). Building this system involved developing probabilistic models of the strengths and weaknesses of humans and computers for different types of tasks such as predicting part locations, attributes, and classes. The system selects automatically which questions to ask human users – the most informative questions are chosen by maximizing an information gain criterion – this reduces on average the amount of interaction that is needed to achieve a satisfactory answer.

A second realization is that it is not possible to build the bird field guide without help from humans. Computer vision researchers dream of building machines that discover structure in images automatically and that can learn visual categories without supervision [9,10]. However, it is not realistic that all the information necessary to build a field guide will be gathered without the help of human experts. Humans have bodies, explore the world, take things apart. There are things that only humans know. Key to Visipedia's success is enabling humans to share their visual knowledge. It is useful

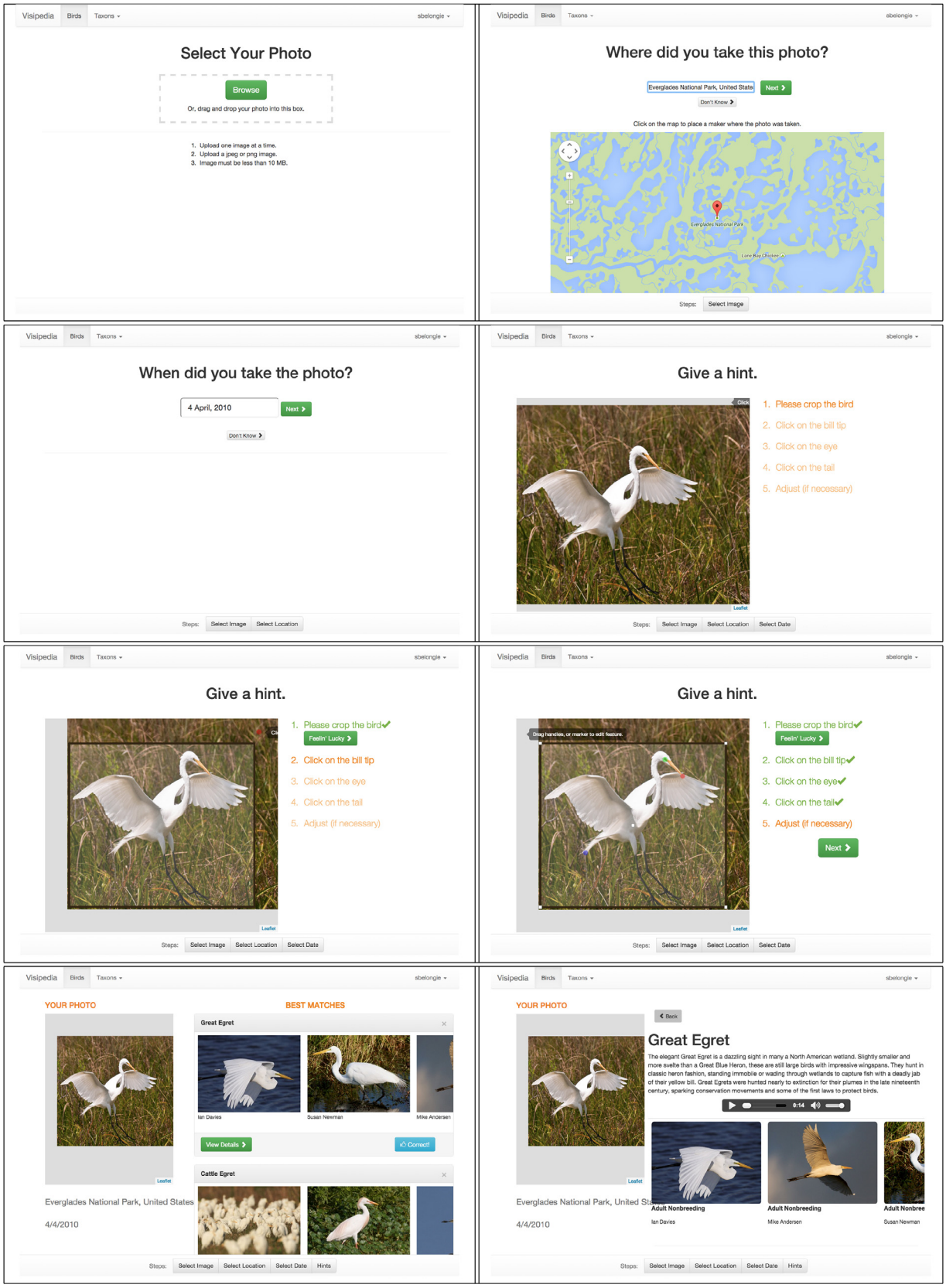

Fig. 3. Illustrating the use of the Merlin to classify the image of a bird (left to right, top to bottom). Merlin may be accessed at http://merlin.allaboutbirds.org/photo-id. The user first uploads a picture (the photograph used in this example was taken from http://parkorbird.flickr.com), then provides the system with the location and date where the photo was taken ('don't know' is a legal input). Finally, the user may click on three landmarks (bill tip, eye and tail end) to help the system locate the bird. The system outputs a list of likely bird species, prioritized by probability. The user may then access additional information by clicking on live links provided with the output.

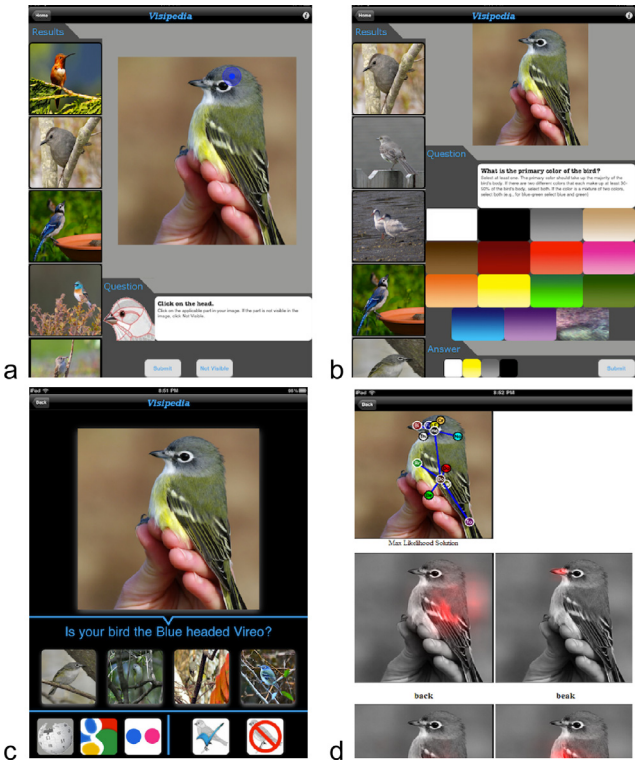

Fig. 4. Screen capture of an iPad app for bird species recognition: A user takes a picture of a bird she wants to recognize, which is uploaded to a server. The server runs computer vision algorithms to localize parts and predict bird species. The computer system intelligently selects a series of questions to ask that are designed to reduce its uncertainty about the predicted bird species as quickly as possible. **(a)** The system poses the question *where is the head?* The user's response is used to refine part location and class probability estimates. **(b)** The system chooses another question: *what is the primary color of the bird?* **(c)** The system thinks that the bird is a *Blue-headed Vireo*. **(d)** Debugging output of the algorithms shows detected part locations and part probability maps.

therefore to explore the steps that were necessary to harvest from humans the expertise and information that was necessary to build and train *Merlin*.

3. Harvesting annotations from experts and non-experts

How are machine vision systems trained? A large and carefully labeled training set is the necessary starting point. Current visual categorization systems [1,2] have order of 10^8 parameters. Regularization, dropout and other techniques are making it feasible to train many parameters with relatively few training examples. However, for best results datasets must contain at least $10^5 - 10^6$ images. Where can all these images be found?

Labeled images of common entry-level categories (birds, bananas, Buddhas) are usually available in vast quantities on the web. A Google Image search for 'bird' will return millions of results. Not all of these will contain birds, but the process of flagging the bad images demands no special expertise of a human annotator. Crowdsourcing services, such as Amazon Mechanical Turk[2] (AMT), that have the ability to recruit a large number of untrained annotators offer an efficient and cost-effective means of cleaning such datasets [11–13]. Thus, the creators of *Park or Bird* could obtain a sufficient number of labeled training images quickly and inexpensively.

Things change when one makes the step from entry level to fine grained categories. To start with, for some species it is difficult to locate suitable training images. Typing "dark eyed junco" into Google Images will yield quite a few nice pictures – nowhere near 10^5 though. Are all the pictures correctly labeled? Can an inexperienced human annotator reliably detect mislabeled images? The answer is, unfortunately, 'no.' A recent study [14] reveals that the fine grained categories in CUB-200 [15] and ImageNet [16], both of which used AMT to clean the datasets, have significant type I and type II errors. While as computer vision researchers we aspire to develop classification models that are robust to such errors, in the context of Visipedia image labels that are 'more or less correct' will not suffice. Visipedia draws upon the talents of experts to certify the labels so that users needn't worry that Black-capped Chickadees and Carolina Chickadees might be mixed up in the shortlist of possible matches.[3] In this respect, Visipedia offers users not only a classification engine but also an expert-curated resource for browsing and learning. The importance of a computer-assisted centralized registry of all species, where experts' opinions may be integrated and where computer vision 'robots' may assist in pointing out inconsistencies and redundancies, is underscored by the recently released World Register of Marine Species (WoRMS) study, which includes the finding that "of the 419,000 species names recorded in the scientific literature, nearly half (190,400) have been shown to be duplicate entries."[4]

Birds is a well studied taxon with few remaining controversies. In building *Merlin* we felt comfortable relying on the expertise of two ornithologists working at the Cornell Lab of Ornithology, who were motivated to collaborate with us. We trust their judgment in building the birds' taxonomy, describing their attributes and labeling species in pictures. However, as we move forward it is important to remember that experts are not necessarily oracles. In fact, Visipedia is potentially most useful in situations where there is no oracle. In Section 8 we describe our first steps towards a no-oracle Visipedia, one where the machine is able to estimate the reliability of a given human annotation and whom to trust.

One more thought: 'expert' does not necessarily denote a card-carrying academic. This may be true in the context of galaxy classification and fine distinction of bird species. However, any human with good eyesight is an expert in the task of detecting a person or a dog in a picture. We found that the workforce in plentiful supply on AMT holds value even for fine grained visual categorization – the study mentioned above [14] found that expert and non-expert annotators perform roughly equally well in part localization tasks (e.g., "click on the eye"). This is good news since training a machine to identify these landmarks is a key step for species classification (Section 2.1).

4. Discovery and taxon formation

In building *Merlin* we assumed that a universally acknowledged set of experts exists, along with well-defined and agreed-upon knowledge about the domain (e.g., ornithology). What if some of those elements did not exist?

Let's think for a moment about the task of identifying a bird. How do we learn to recognize a bird? Where does the primary information come from? Let's say that an ornithologist at some point has 'discovered' a new bird species in the field and described it in detail. Her descriptions are based upon a consensus of the distinguishing characteristics of birds and an agreement on the overall taxonomy. What if such a tradition or consensus had not existed? Imagine for a moment that no bird had ever been seen. What would it be like to find the first specimens, reason about their similarities and differences, realize that they may be grouped into thousands of species, and organize them into a taxonomy?

[2] https://www.mturk.com.

[3] http://www.sibleyguides.com/bird-info/black-capped-chickadee/black-capped-carolina-chickadee.

[4] http://www.bbc.com/news/science-environment-31851525 and http://www.marinespecies.org.

These are not idle questions. Most human activities, from astronomy and architecture, to fashion, pathology and zoology, involve creation and discovery of new categories, and organizing principles. *'Folksonomies'* [17] regularly emerge, even outside science, to help people organize and exchange information efficiently. Description and taxonomization of new taxa and corpora are as important as annotation and knowledge dissemination [18].

We would like to draw upon computer vision and machine learning to help in this process. While it is exciting to think about autonomous machines embarking in discovery missions (e.g., on distant asteroids and in the depths of the sea), in most areas machines are not about to substitute for humans – humans are still the primary agents of discovery, and humans end up holding much of the knowledge. Machines need to be able to learn from humans, help humans organize the knowledge they have acquired, assist humans in navigating and making sense of an ever increasing volume of pictures and other data.

5. Human contributors to Visipedia

If we wish to build a system that is able to harvest useful information from people, it is useful to pause for a moment and reflect on the fact that individuals have a diverse palette of goals and levels of expertise. In the course of developing *Merlin* we identified at least five different types:

Experts – Professionals who focus their work on discovery and teaching in a given domain. They may be able to answer difficult questions. They are few in number, their time is limited and expensive. They may disagree spectacularly and not be aware of it.

Citizen scientists – Passionate hobbyists: avid bird watchers, amateur astronomers, fashionistas. They dedicate energy and attention to their favorite subject, can answer many questions, will collect and annotate valuable data. They are often organized by levels of expertise. They are keen to improve their knowledge and communicate it to others.

Annotators – Workers who, given minimal training, can carry out general tasks such as drawing bounding boxes, clicking on parts and labeling attributes. Their reward is a combination of money and fun. They may be quickly recruited in large numbers through online services such as AMT.

Computer scientists – This refers to us, the pattern recognition, machine learning and computer vision researchers. We are keen to help by building intelligent machines, but we do not know much about the specifics of the thousands of taxa and fields of expertise.

Users – Members of the general public who would like to have their visual queries answered. They are often willing to work a little in order to obtain useful information [7].

These people do not necessarily know each other, and may only be communicating with each other by means of the information that is exchanged on the topic of mutual interest. Moreover, initially they are not known to the machine(s) either. A network of trust must somehow come into being between multiple people and machines.

One may think of these groups of people as different 'resources' which may help Visipedia if they are properly instructed, paid, entertained, networked. A systematic study of how best to harness these resources in a single system has barely begun [14].

The definitions above are useful in the context of knowledge that is not widely available. As noted in the previous section, for simple tasks, such as finding dogs in pictures, everyone with good eyesight is an expert.

6. Components of Visipedia

Our discussion so far helps us bring into focus a number of components of Visipedia, comprising humans and machines:

Software system – Composed of (a) a back-end database where pictures, annotations, links, taxonomies and other useful pieces of information are collected and periodically updated, (b) Graphical User Interfaces that are designed to crowdsource image labeling, design taxonomies, etc., (c) web crawlers to collect and organize pictures and related annotations on-line, (d) an 'operating system' managing and coordinating the available resources.

Images – The primary objects of interest. These come from different sources: they are uploaded by the members of a community, crawled on-line, acquired by scientific instruments.

Taxon manager – A person, or a piece of software, who minds a taxon, i.e., a domain of knowledge (e.g. 'birds' or 'antique furniture'). The taxon manager has to locate the experts and interact with them to obtain information about the parts, attributes and taxonomy. It then has to decide how to best annotate the available images, by a combination of experts, annotators and machines. It has to manage the social network of people who care about the taxon and contribute to it, establish levels of 'reputation' and degrees of access to the data.

People – A community of experts, citizen scientists, users, etc., as discussed in Section 5.

Inference engines– Computer Vision and Machine Learning algorithms to analyze images, recognize objects and their parts, classify species, taxonomize, measure attributes, link together related images and image patches. Some of these processes are designed to ask engaging and informative questions from human observers. Algorithms and machines that crunch on human-supplied and machine-supplied labels and features in order to estimate the labels of the images in the database, confidence intervals and other useful knowledge.

We envision that the process of bringing any given taxon 'online' on Visipedia will necessitate the appointment, adaptation and/or onboarding of the above components, as appropriate for the taxon.

7. Questions

Our discussion so far raises a number of questions. Answering these questions satisfactorily is all the more important if one wishes to automate Visipedia.

1. Who are the 'experts' of a certain domain, and who gets to decide who is an expert? Since this leads to an infinite regression, can a self-organizing, decentralized process do this? What happens when experts do not agree?
2. How can the specific expertise of each human be assessed? What are the main dimensions of expertise?
3. How can the system assess the reliability of the information that is provided by annotators, citizen scientists and experts? Can the distributed system estimate the degree of confidence that a given piece of information is correct, e.g., that the bird in a picture is a snowy egret?
4. How can the 'taxon meta information' (taxonomy, parts, attributes) be obtained and organized, and can this be done automatically? What are the steps through which a taxon comes to life, with a community of people around it?

We explore these questions in the next sections.

8. Who will judge the judges?

Can Visipedia automatically assess the accuracy of experts and annotators? This is a chicken-and-egg problem. If we had reliably annotated images, then we could measure the accuracy of human annotators. Conversely, if we had accurate human annotators, we would be able to obtain reliably annotated images. However, at the beginning the images have not been annotated and the system does not yet know who are the experts of a given domain.

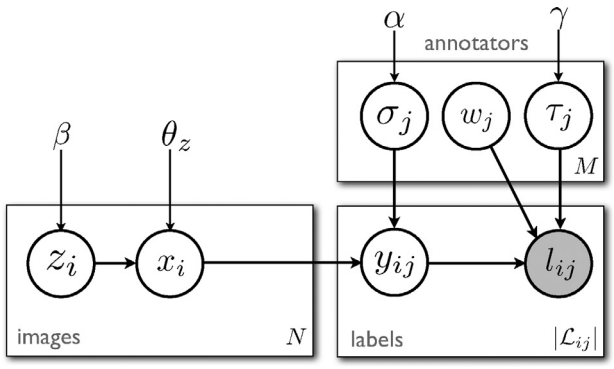

Fig. 5. Model of the annotation process, adapted from [19]. The index i refers to images and the index j refers to annotators (which may be anything between naive and expert). The task is binary annotation, e.g. 'does the bird in this image belong to species X ?' (see Fig. 8). The variable z_i indicates the unknown ground truth ($z_i = 1$ means that the bird in image i does belong to species X). The variable x_i indicates a vector of relevant measurements (e.g. color of plumage, length of beak etc.) carried out by the *ideal observer*, i.e. the best birder. The vector y_{ij} indicates equivalent measurements as carried out by observer j; this set of measurements is modeled as the ideal measurements x_i corrupted by 'noise' that is specific to observer j and is parametrized by σ_j. The larger the noise, the more incompentent the observer. Observer j will label image i with a binary label l_{ij} based on the measurements y_{ij}. In order to do so, the observer will use an observer-specific classifier parameterized by w_j, and an observer specific bias t_j. Only the labels l_{ij} are observable. The rest of the quantities are estimated from the labels thanks to the model.

A practical solution to this problem is appointing a human 'editor' who will identify experts who can, and will, annotate a well defined taxon (e.g., 'birds'). This is similar to starting a new scientific journal. Natural selection makes sure that only the fit survive. If the editor is skillful, then the taxon will be annotated correctly and it will attract many users. If, instead, the editor is inept, the annotations will be poor and the community will not thrive. Thus, random initialization and natural selection appears to be a viable solution, if a bit wasteful.

It is perhaps more interesting to explore an alternative solution that requires no explicit leadership. Would it be possible for the 'truth' to emerge as a result of the annotations that are supplied by many people? Let's suppose for a moment that experts will mostly agree, while incompetent annotators will produce inconsistent annotations. In this case if every image is annotated by multiple people, it should be possible to see a pattern emerge: those annotations that are in agreement will be deemed more reliable. Those annotators who tend to agree will be deemed to be the 'experts.' Will this work?

The answer is a tentative 'yes.' We explored this issue in the context of binary bird classification as well as a number of synthetic binary labeling tasks. We found that, if one is willing to model the process of image annotation, one may jointly estimate the likelihood of the image's class, the reliability of human classifiers, as well as the difficulty of classifying a specific image [19]. Our experiments show that for difficult tasks it is advantageous to model human annotators in some detail. Our model included three characteristics (see Fig. 5): the *incompetence* (σ) of the human annotator (modeled as a noise source), the *bias* (τ) of the human annotator (some are more conservative and require more evidence than others in order to say 'yes') and the *classification strategy* (w), i.e., the parameters of the classifier.

Allowing the model to assign a different classifier to each annotator captures the fact that different people may use different features and classification criteria in order to accomplish the same visual task. Consider human annotators who are asked to discriminate between pictures of seagulls and pictures of hawks. Some may use the fact that seagulls tend to be white-gray or light brown, while hawks tend to be dark brown. Others may use the fact that seagulls have webbed feet, while hawks have talons. A third group may use the bill shape. A fourth group may use a combination of these criteria. Each group will perform equally well on good quality images. However, in a picture showing only the head of the bird annotators who look at the feet will be at a disadvantage. Similarly, a photograph that shows the bird silhouetted against the sun will present difficulty for annotators who rely on bird color. Our experiments show that when a sufficient number of labels are collected from a sufficient number of annotators, modern inference techniques allow the system to estimate the annotators' internal parameters, as well as each image's most likely label, and its intrinsic difficulty [19].

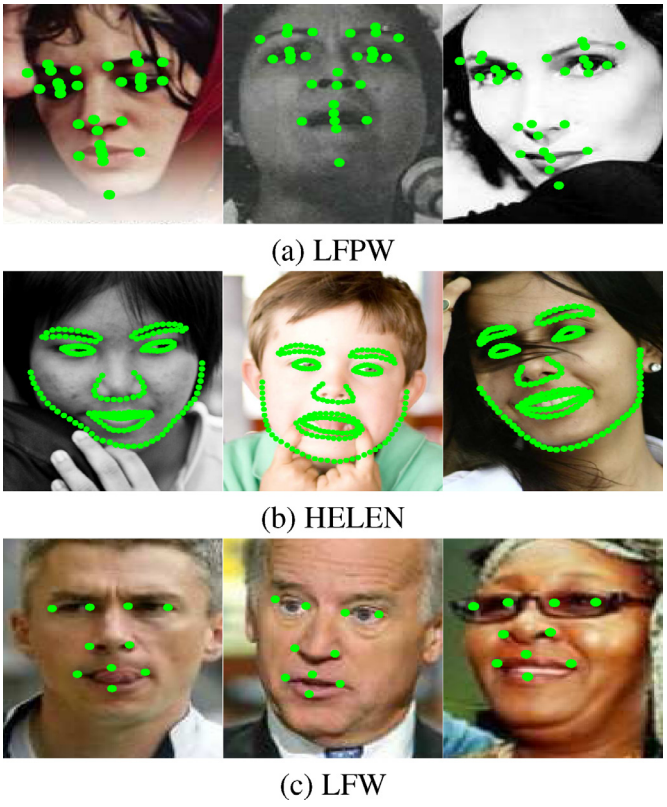

Fig. 6. Experts often develop different 'partonomies', i.e. subdivisions of a given category into parts; this poses a challenge for Visipedia. Here facial landmarks in different face recognition datasets are shown (source: [20]). See also Fig. 7.

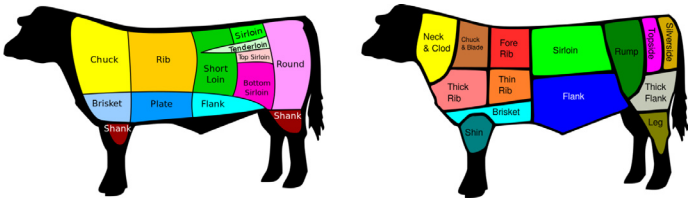

Fig. 7. Example of a 'folksonomy' [17] : cuts of meat in the U.S. and in Britain (source: Wikipedia).

These initial experiments tell us that it is possible to estimate which annotators (or experts) we can trust and, simultaneously, which is the likely interpretation of the image, and which images are too difficult to annotate and should be discarded. The chicken-and-egg problem may be somehow circumvented. Furthermore, it is possible to measure other characteristics of the annotators and experts, such as their level of risk aversion and which features are they likely using to accomplish their task. These parameters are useful to integrate the experts' labels into a final estimate. Much remains to be done, but all this is encouraging.

Sometimes the annotators' parameters are clustered into groups. We will discuss this useful observation in Section 9.

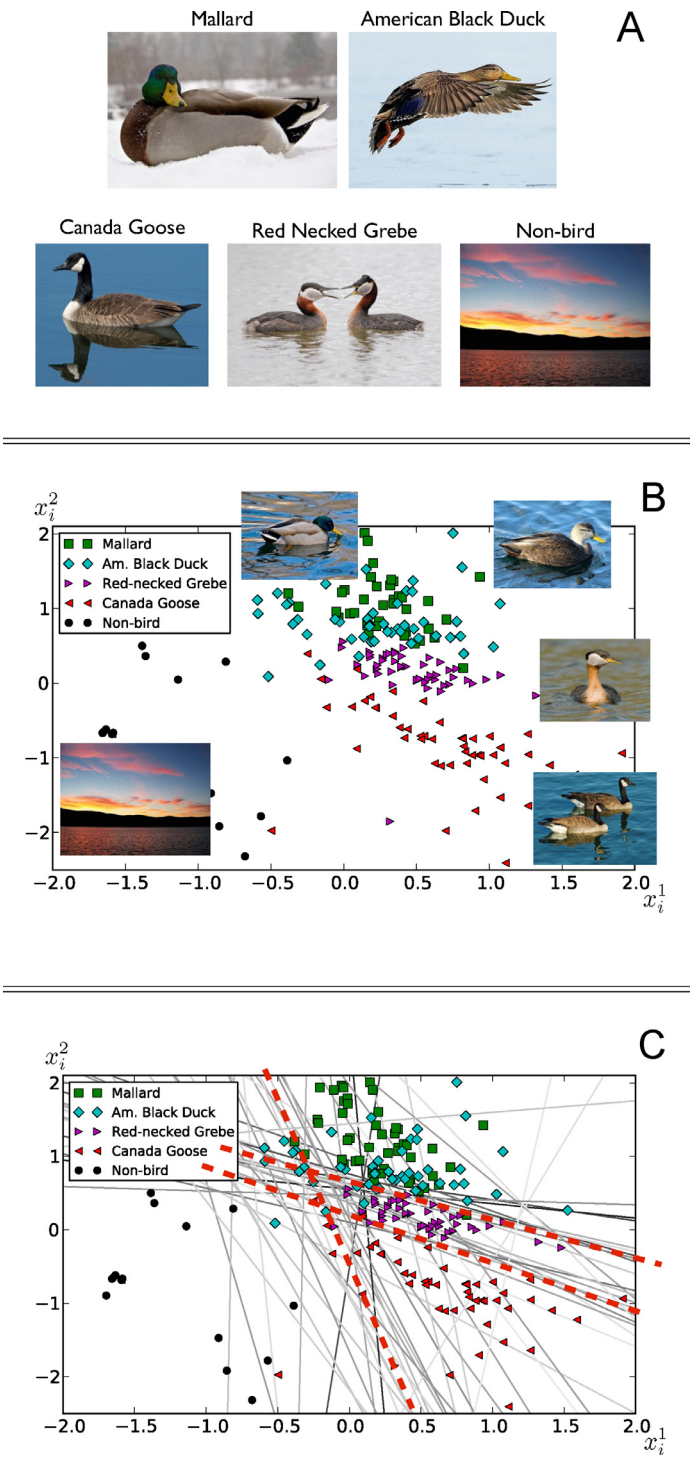

Fig. 8. Fifty annotators were asked to label one hundred images as to the presence of ducks. (A) Some images showed ducks (mallards, and American black ducks), some geese, some grebes, and some only showed birdless wetlands. (B) An inference algorithm that was based on the model shown in Fig. 5 estimated the 'latent coordinates' x_i of each image (one point per image). These coordinates are an estimate of the the two most informative measurements that an ideal observed would carry out. Ducks, grebes, geese and birdless images are separated from each other. (C) Each line indicates an estimate of the linear classifier (w_j, τ_j) that was used by each observer. The observers may be divided into three distinct groups: those who correctly separate the ducks from the rest of the images, those who group the grebes with the ducks and those who label as 'duck' any water bird. This suggests that it is possible to discover different schools of thought amongst the annotators. (Adapted from [22].). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Fig. 9. Top: Questions of the form "Is object *a* more similar to *b* than to *c*?" offer a useful way to collect similarity data from crowd workers. Bottom: Two dimensional embedding of food images computed using this triplet-based similarity data.

9. Schools of thought

We have mentioned before that experts may disagree. This is true both in the way they organize knowledge and in the way they classify data.

It is not infrequent to observe that different experts develop different *partonomies*, i.e., different organizations of a category into parts and attributes (see systems for face annotation in Fig. 6). Folksonomies [17] may also differ (see cuts of meat in Fig. 7). This poses a challenge for Visipedia: how is an information system supposed to choose between different 'dialects' that have emerged to describe the structure of a category? We believe that the best solution is to let different sub-communities use their respective preferred conventions and develop an automated translation system to reconcile information that is collected according to different systematizations [21].

We observed the second phenomenon, different classifications, both in a synthetic experiment and in an experiment where a number of annotators were asked to identify 'ducks' in pictures [19,22]. Most pictures contained water birds (two species of ducks, one species of grebes and one of geese) and some contained waterscapes with no birds. Annotators may be clustered into three groups (see thick dashed red lines in Fig 8 (bottom)). The first group carried out the task correctly, classifying the two species of ducks as 'duck'. A second group recognized ducks and grebes as 'duck'. A third group labeled every bird as 'duck', including geese and grebes.

These groups may be thought as *'schools of thought,'* to use a somewhat grandiose term. While it is impossible, just from the data, to know whether a school of thought is correct, it is possible to detect the presence of three distinct points of view. The observation that there may be different schools of thought refines the assumption that was made in Section 8, namely that the best annotators will tend to

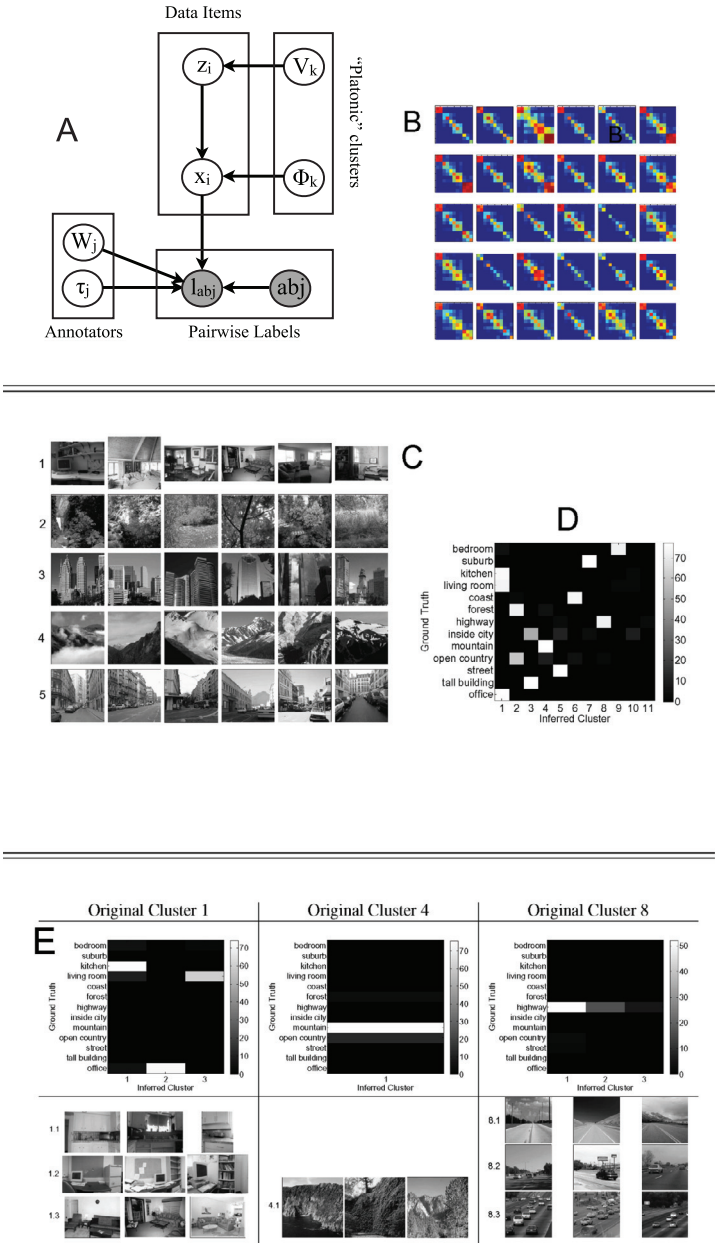

Fig. 10. Crowdsourced clustering ('crowdclustering') [36]. (A) Probabilistic model of the labeling process. Each annotator j has a different metric W_j that they use for clustering. The metric is not known in advance, but may be inferred from the data using the model. The rest of the model is similar to that shown in Fig. 5. (B) If a ground-truth clustering is known, the labels produced by each annotator may be summarized by a 'confusion matrix'. Each annotator follows idiosyncratic criteria and produces a different confusion matrix. (C) Images used in a clustering experiment. The images belong to the 15-scenes dataset [37]. (D) Clustering obtained in the first pass. Scenes of 'kitchen', 'living room' and 'office' were clustered together into cluster 1. (E-left) Re-clustering images belonging to cluster 1 produces three further clusters, corresponding to different rooms. (E-center) Re-clustering images belonging to cluster 4 ('mountain') produces no further clustering. (E-right) Re-clustering images belonging to cluster 8 refines 'highway' into three clusters, one with high traffic density, one with low traffic density and one with no vehicle in sight. Thus, recursively re-clustering each cluster will produce a taxonomy that best fits the crowd's perception of the data.

agree, while mistaken annotators will disagree. The world is not that simple.

We postulate that detecting 'schools of thought' amongst experts is beneficial. In a number of occasions we have observed that experts are not even aware that they disagree. Making disagreement explicit, and recognizing the existence of different schools of thought, may help experts come to terms with the source of their disagreement and understand whether additional data need to be collected and experiments performed to resolve the ambiguity, or whether their differences are purely the result of convention.

10. Category and attribute discovery

The bird taxon is composed of distinct species crisply organized in a taxonomy. But what about the universes of things that are not so neatly delineated, or for which a taxonomy may be more the result of qualitative judgment and convention? While taxa such as *handbags* or *food* possess visual complexity (i.e., part and attribute decomposition) comparable to that of birds, there may be no underlying inherent taxonomy. The 'perfect' taxonomy, however, may not be as vital as for birds – potential users of a Visipedia for such taxa might be

satisfied with loose clusterings as well as relationships [23], e.g. to help browse the data, rather than in a sharp classification answer.

Will some crowdsourcing method help discover how people organize visual information within a taxon? The answer depends on the assumptions that one makes on the nature of such an organization, for example whether the most natural representation is continuous or discrete.

Suppose for a moment that people think of a given taxon as a continuum in a suitable space, where individual items are points, e.g., the facial physiognomy of caucasian people, or the color of different cloths. Psychologists interested in this question developed methods to probe our perceptual system and estimate low-dimension embeddings of the relevant stimuli [24,25]. Recently, machine learning researchers have become interested in this question as well [26]. Low-dimensional embeddings that capture perceptual similarity between images may be obtained, for instance, from three-way comparisons between images that are carried out by annotators or users who interact with the system. Fig. 9 shows an example of an embedding of food from the Yummly dataset [27]. This embedding was obtained by learning a *crowd kernel* [28] from batches of triplet-style questions of the form "does food *A* look like it tastes more like food *B* or food *C*?" This experiment in crowd kernel learning did not make use of automatically computed image features, and it would be interesting to explore ways to accelerate the creation of the embeddings through a combination of human input and machine vision.

Let's now take the discrete point of view. Are classes, taxonomies and attributes directly discoverable from image collections? Pattern recognition has long recognized that 'clustering' and 'grouping' are fundamental and difficult questions. In the case of vision, the natural features are the pixels – not a good starting point for clustering. The relevant features (beak shape, color of plumage) are not directly readable from the data – they have to be inferred from the data, which compounds the problem. Computer vision researchers have identified this fascinating and challenging question, and the literature on the topic is quite rich [9,29–31]. While a number of approaches look promising, no system works well yet. What about involving human observers? Interesting approaches to inferring parts and attributes with human help have been proposed by [32–35].

What about the process of inferring classes and taxonomies? Can the task be crowdsourced, so that machines may learn from the 'wisdom of the crowd'? Perhaps untrained annotators may be able to detect similarities and differences between objects in images, even if they cannot verbalize them. These annotations may be later combined automatically to infer a coarse taxonomization, which may be later refined by experts. There are two potential difficulties: (a) When the dataset is sufficiently large no human has the ability to explore it all and draw conclusions, the memory of an individual is not sufficient for carrying out a large unsupervised clustering task. Realistically, each annotator will only access a small subset of the images and will work on clustering within that manageable subset. Solving smaller tasks deprives the annotators of the broader context, and therefore the clusterings that are proposed on small subsets may be meaningless in the larger context. (b) If nobody has yet seen the data, there may not be universally acknowledged criteria, and each annotator may develop different criteria for clustering the data.

We found empirically that these obstacles may be overcome if the proper probabilistic model is used to account for the annotators' differences [36]. We call *crowdclustering* the idea of aggregating automatically many potentially inconsistent hand-made clusterings. A method we proposed recently to tackle this challenge is illustrated in Fig. 10. In this setting a large number of annotators solve small image clustering problems making use of an ad-hoc graphical user interface (GUI) where a panel containing a 6×6 grid of images is shown. Each user facing a clustering problem defines an arbitrary number of clusters and associates each image to a cluster (a 'no cluster' option is also available). For each pair of images a, b on the panel this produces a binary label l_{abj} indicating whether annotator j believes that images a, b belong to the same cluster ($l_{abj} = 1$) or not ($l_{abj} = 0$). While each annotator has different views on the best clustering criteria (see confusion matrices in Fig. 10 (B)), and only has a local view of the data, the set of labels obtained from all annotators may be combined by a suitable maximum likelihood algorithm (the probabilistic model is shown in Fig. 10 (A)) to obtain a global clustering of the data and thus produce an interpretation of the data as a set of 'classes' (see Fig. 10 (C,D)). A taxonomy may be obtained by repeating the clustering operation on the images belonging to clusters obtained in the previous step. At each step annotators have a tighter context and thus refine their clustering criteria (see Fig. 10 (E)). This recursive clustering process produces the taxonomy that best fits the ensemble of the annotator's perception.

To summarize this section: a number of exploratory studies shows that it is possible for machines to discover the categorization, the taxonomy, the parts and the attributes of objects in a taxon. The information is in part present in images and in part it is present in the head of experts. Sometimes the information is distributed across a crowd. By using appropriate models and inference techniques the information may be collected and pieced together. Of curse, much more work is needed in order to build fully automated systems that can deal with any situation by probing the most informative sources (whether the data, experts or annotators) and using the most appropriate models and inference algorithms.

11. Discussion and conclusion

'Visipedia' is a system for capturing, accumulating, organizing and distilling visual knowledge, akin to what Wikipedia is for verbal knowledge. Visipedia will be able to harvest information from experts, become more accurate with use, and offer visual expertise to anyone, anywhere, at any time.

In exploring the likely building blocks and implications of Visipedia we have touched upon a number of themes that we think deserve the attention of researchers in computer vision and machine learning. Some of these themes extend outside the traditional scope of these disciplines.

The first is a change of perspective. The 'vision system' is not a 'box' to be designed, trained and handed over to a customer. Rather, it is an organic, distributed system composed of machines and people that evolves over time. Within this system the machines learn from people, and people from machines. Machines and people are often synergetic in accomplishing a task. Images are transformed from 'data' into 'information' in the process. This broader definition of the system changes substantially both theory and practice. For example, the key assumption that training sets are annotated by an infallible oracle is not true any longer. Understanding the role that different groups of people play, and modeling their goals and expertise, is as important as developing good computer vision and machine learning algorithms.

The second is that visual knowledge is often implicit in the head of experts, rather than directly available. One has to find ways to harvest knowledge from people, and transfer it into machines in a machine-friendly form. Furthermore, experts may disagree with each other. Making disagreement explicit and modeling it, e.g., as different 'schools of thought,' is important – both to discover which knowledge is available and to help humans improve their knowledge through discussion, hypothesis testing and experimentation.

Third, the converse of harvesting visual expertise is teaching and training. Citizen scientists, annotators and even experts are rewarded by the prospect of learning and honing their skills. If the system can model visual expertise, it is also able to offer a human trainee the most useful next piece of knowledge to improve his or her expertise [38]. This, in turn, will make the system more capable.

The fourth theme is computer-aided exploration and discovery in image sets that are too large for humans and for which humans have not yet formed a good model. The explorer – a scientist, a journalist, an urban planner – is helped by a machine that learns to carry out an increasing number of tasks as new categories are discovered and as new concepts are developed. Conversely, we may think of humans as a means to achieve human-assisted machine learning: the machine asks intelligent questions when it is confused, and knows whom to ask which question in order to obtain information efficiently. These are two sides of the same coin.

A fifth theme is that some tasks are too difficult for machines and too large for a single human. Sometimes knowledge is distributed over human communities. Capturing distributed knowledge and making use of crowdsourcing to solve big image exploration tasks is very useful and appears to be feasible. Examples of such tasks include capturing folksonomies in diverse fields such as fashion and medicine, as well as clustering and organizing large bodies of data.

Thinking about Visipedia is pushing us beyond the traditional scope of pattern recognition, computer vision and machine learning. We are invited to think about vision and learning systems as dynamical systems composed of data, machines and humans. We believe that it is well worth embracing the challenge.

Acknowledgments

We dedicate this article to Jitendra Malik, our thesis adviser and Jedi Master, on the occasion of his being awarded the International Association for Pattern Recognition King-Sun Fu Prize. Jitendra's ideas on scene understanding have inspired much of our work. His energetic advocacy of deeper thinking in computer vision has often pushed us to ask more interesting questions. We are, in turn, grateful to our students and post-docs, the new generations of Jitendra's academic family, who helped us shape our thoughts on Visipedia during the past five years: Ron Appel, Steve Branson, Ryan Gomes, Grant van Horn, Catherine Wah, Peter Welinder, Michael Wilber.

Steve Branson, Kristen Grauman, David Hall, Grant van Horn and two anonymous referees read our manuscript carefully and provided very valuable advice to make it more readable, consistent and better organized.

We gratefully acknowledge funding from Google, from an ARO/JPL-NASA Stennis Grant NAS7.03001 and the ONR MURI Grant N00014-10-1-0933.

References

[1] A. Krizhevsky, I. Sutskever, G. Hinton, ImageNet classification with deep convolutional neural networks, in: Proceedings of the Conference on Neural Information Processing Systems (NIPS) 25, 2012, pp. 1106–1114.
[2] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, T. Darrell, Caffe: Convolutional architecture for fast feature embedding, arXiv preprint arXiv:1408.5093 (2014).
[3] E. Rosch, Principles of categorization, in: E. Rosch, B.B. Lloyd (Eds.), Cognition and categorization, Erlbaum, Hillsdale, NJ, 1978.
[4] S. Branson, G. van Horn, S. Belongie, P. Perona, Bird species categorization using pose normalized deep convolutional nets, Arxiv preprint arXiv:1406.2952 (2014).
[5] Y. Taigman, M. Yang, M. Ranzato, L. Wolf, DeepFace: Closing the gap to human-level performance in face verification, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014.
[6] N. Zhang, J. Donahue, R. Girshick, T. Darrell, Part-based R-CNNS for fine-grained category detection, in: Computer Vision–ECCV 2014, Springer, 2014, pp. 834–849.
[7] S. Branson, C. Wah, B. Babenko, F. Schroff, P. Welinder, P. Perona, S. Belongie, Visual recognition with humans in the loop, in: Proceedings of the Eleventh European Conference on Computer Vision (ECCV), Heraklion, Crete, 2010. http://www.vision.caltech.edu/visipedia.
[8] S. Branson, G. Van Horn, C. Wah, P. Perona, S. Belongie, The ignorant led by the blind: a hybrid human–machine vision system for fine-grained categorization, Int. J. Comput. Vis. 108 (1-2) (2014) 3–29.
[9] M. Weber, M. Welling, P. Perona, Towards automatic discovery of object categories, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Hilton Head Island, USA, 2000.
[10] R. Fergus, P. Perona, A. Zisserman, Object class recognition by unsupervised scale-invariant learning, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2, 2003, p. 264.
[11] L. Fei-Fei, R. Fergus, P. Perona, Learning generative visual models from few training examples: An incremental Bayesian approach tested on 101 object categories, in: Proceedings of the IEEE CVPR Workshop of Generative Model Based Vision (WGMBV), 2004.
[12] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, L. Fei-Fei, ImageNet: A large-scale hierarchical image database, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2009.
[13] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, C.L. Zitnick, Microsoft COCO: Common objects in context, in: Computer Vision–ECCV 2014, Springer, 2014, pp. 740–755.
[14] G. Van Horn, S. Branson, R. Farrell, S. Haber, J. Barry, P. Ipeirotis, P. Perona, S. Belongie, Building a bird recognition app and large scale dataset with citizen scientists: The fine print in fine-grained dataset collection, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 595–604.
[15] P. Welinder, S. Branson, T. Mita, C. wah, F. Schroff, S. Belongie, P. Perona, Caltech-UCSD Birds 200, California Institute of Technology. CNS Technical Report 2010-001, 2010.
[16] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, L. Fei-Fei, ImageNet: A large-scale hierarchical image database, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2009, pp. 248–255.
[17] T. Vander Wal, Folksonomy, online posting, Feb 7, (2007). http://vanderwal.net/folksonomy.html.
[18] M. Brescia, S. Cavuoti, G.S. Djorgovski, C. Donalek, G. Longo, M. Paolillo, Extracting knowledge from massive astronomical data sets, in: Astrostatistics and Data Mining, Springer, 2012, pp. 31–45.
[19] P. Welinder, S. Branson, S. Belongie, P. Perona, The multidimensional wisdom of crowds, in: Proceedings of the Conference on Neural Information Processing Systems (NIPS), 2010.
[20] X.P. Burgos-Artizzu, P. Perona, P. Dollár, Robust face landmark estimation under occlusion, in: Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2013.
[21] S. Branson, K.E. Hjorleifsson, P. Perona, Active annotation translation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2014, pp. 3702–3709.
[22] P. Welinder, Hybrid human-machine vision systems: Image annotation using crowds, experts and machines, California Institute of Technology, 2012. (Ph.D. thesis)
[23] T. Malisiewicz, A. Efros, et al., Recognition by association via learning per-exemplar distances, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2008, pp. 1–8.
[24] J.B. Kruskal, Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis, Psychometrika 29 (1) (1964) 1–27.
[25] A. Tversky, I. Gati, Similarity, separability, and the triangle inequality., Psychol. Rev. 89 (2) (1982) 123.
[26] M. Schultz, T. Joachims, Learning a distance metric from relative comparisons, in: Proceedings of the Conference on Advances in neural information processing systems (NIPS), 2004, p. 41.
[27] M.J. Wilber, I.S. Kwak, S.J. Belongie, Cost-effective hits for relative similarity comparisons, in: Proceedings of the Conference on Human Computation & Crowdsourcing (HCOMP), 2014.
[28] O. Tamuz, C. Liu, S. Belongie, O. Shamir, A.T. Kalai, Adaptively learning the crowd kernel, in: Proceedings of the International Conference on Machine Learning (ICML), 2011.
[29] K. Grauman, T. Darrell, Unsupervised learning of categories from sets of partially matching image features, in: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 1, IEEE, 2006, pp. 19–25.
[30] J. Sivic, B.C. Russell, A. Zisserman, W.T. Freeman, A.A. Efros, Unsupervised discovery of visual object class hierarchies, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2008, pp. 1–8.
[31] E. Bart, I. Porteous, P. Perona, M. Welling, Unsupervised learning of visual taxonomies, in: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), 2008.
[32] O. Russakovsky, L. Fei-Fei, Attribute learning in large-scale datasets, in: ECCV Workshops, Springer, 2010, pp. 1–14.
[33] D. Parikh, K. Grauman, Interactively building a discriminative vocabulary of nameable attributes, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2011, pp. 1681–1688.
[34] A. Kovashka, K. Grauman, Attribute adaptation for personalized image search, in: Proceedings of the IEEE International Conference on Computer Vision (ICCV), IEEE, 2013, pp. 3432–3439.
[35] P. O'Donovan, J. Lībeks, A. Agarwala, A. Hertzmann, Exploratory font selection using crowdsourced attributes, ACM Trans. Graph. (TOG) 33 (4) (2014) 92.
[36] R. Gomes, P. Welinder, A. Krause, P. Perona, Crowdclustering, in: Proceedings of the Conference on Neural Information Processing Systems (NIPS), 2011.
[37] L. Fei-Fei, P. Perona, A bayesian hierarchical model for learning natural scene categories, in: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), IEEE Computer Society, 2005, pp. 524–531.
[38] E. Johns, O. Mac Aodha, G.J. Brostow, Becoming the expert-interactive multi-class machine teaching, arXiv preprint arXiv:1504.07575 (2015).