8-2012

A non-parametric visual-sense model of images: Extending the cluster hypothesis beyond text

Kong-Wah WAN

Ah-hwee TAN
Singapore Management University, ahtan@smu.edu.sg

Joo-Hwee LIM

Liang-Tien CHIA

Citation

WAN, Kong-Wah; TAN, Ah-hwee; LIM, Joo-Hwee; and CHIA, Liang-Tien. A non-parametric visual-sense model of images: Extending the cluster hypothesis beyond text. (2012). *Multimedia Tools and Applications*. 56, (3), 509-534.
Available at: https://ink.library.smu.edu.sg/sis_research/5204

Multimed Tools Appl (2012) 56:509–534
DOI 10.1007/s11042-010-0615-y

A non-parametric visual-sense model of images—extending the cluster hypothesis beyond text

Kong-Wah Wan · Ah-Hwee Tan · Joo-Hwee Lim · Liang-Tien Chia

Published online: 27 October 2010

Abstract The main challenge of a search engine is to find information that are relevant and appropriate. However, this can become difficult when queries are issued using ambiguous words. Rijsbergen first hypothesized a clustering approach for web pages wherein closely associated pages are treated as a semantic group with the same relevance to the query (Rijsbergen 1979). In this paper, we extend Rijsbergen's cluster hypothesis to multimedia content such as images. Given a user query, the polysemy in the return image set is related to the many possible meanings of the query. We develop a method to cluster the polysemous images into their semantic categories. The resulting clusters can be seen as the *visual* senses of the query, which collectively embody the visual interpretations of the query. At the heart of our method is a non-parametric Bayesian approach that exploits the complementary text and visual information of images for semantic clustering. Latent structures of polysemous images are mined using the Hierarchical Dirichlet Process (HDP). HDP is a non-parametric Bayesian model that represents images using a mixture of components. The main advantage of our model is that the number of mixture components is not fixed a priori, but is determined during the posterior inference

K.-W. Wan (✉) · J.-H. Lim
Institute for Infocomm Research, 1 Fusionopolis Way, Singapore, Singapore
e-mail: kongwah@i2r.a-star.edu.sg

J.-H. Lim
e-mail: joohwee@i2r.a-star.edu.sg

A.-H. Tan · L.-T. Chia
School of Computer Engineering, Nanyang Technological University,
Singapore, Singapore

A.-H. Tan
e-mail: asahtan@ntu.edu.sg

L.-T. Chia
e-mail: asltchia@ntu.edu.sg

process. This allows our model to grow with the level of polysemy (and visual diversity) of images. The same set of components is used to model all images, with only the mixture weights varying amongst images. Evaluation results on a large collection of web images show the efficacy of our approach.

Keywords Hierarchical Dirichlet Process · Non-parametric models · Image clustering · Sense disambiguation

1 Introduction

While there is undoubtedly an enormous amount of information on the Internet, their utility is only as good as their accessibility. The past decade has seen development of robust web page ranking methods that harness both intrinsic page information (such as the links, anchor texts, etc) and also extrinsic related information (such as user click-through data). Nonetheless, a major challenge remains when there is inherent ambiguity in the search query. This situation often arises in practice because user queries are generally short and imprecise, and hence may represent many different information needs. For example, the query "jaguar" can refer to the animal or the car.

A common approach to handle query ambiguity is through automatic clustering of web search results. The theoretical underpinning for this approach is based on Rijsbergen's cluster hypothesis [29] that states that the associations between documents convey information about the relevance of documents to requests. A key application of the hypothesis is that documents that are relevant to the query are usually clustered together, and different clusters embody different relevance to the query. In particular, a search engine may assume that the different information needs of an ambiguous query are captured by the topical clusters of the search results.

The validity of Rijsbergen's cluster hypothesis has been borned out by its many successful applications in web search result clustering [37, 44], cluster-topic-based document retrieval models [41], and exploratory browsing interfaces [11]. However, the above successes are mainly limited to web *text* document search, and have not been replicated to image and video search. For example, on the ambiguous query "apple" which can refer to the fruit, the company or the product, most top image search engines[1] return results that haphazardly alternate between the first two meanings. Hence, a user looking for images of iMac would likely be disappointed. In the absence of any disambiguating information on the search query, the better approach, exemplifying the cluster hypothesis, is to present users with clusters of images embodying the multiple dominant interpretations of the original query. The resulting clusters can then be seen as the *visual* senses of the query (see Fig. 1).

Clustering images into their visual senses is akin to *semantic* clustering of the images, which remains a challenging problem. A common approach is to assume that visually similar images are close to each other in feature space, and construct models in which features are generated from a mixture of probability distributions,

[1] We tried Google, Yahoo and Microsoft Bing Image Search.

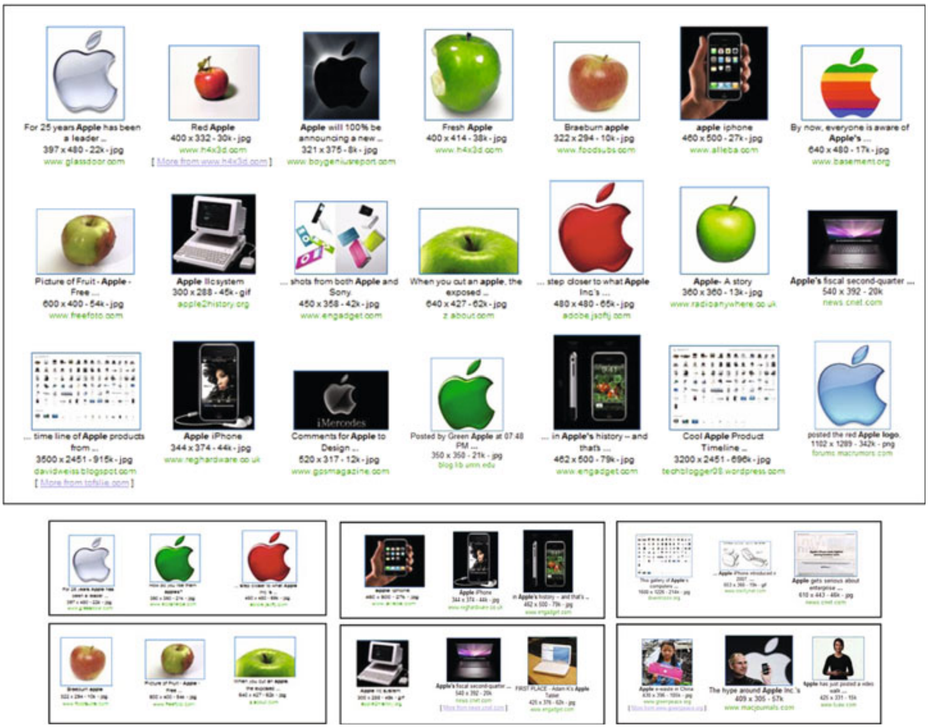

Fig. 1 *Top* Top-ranked "apple" images fom Google Image; *bottom* six possible visual senses (clusters): *logo, fruit, iphone, iMac, drawings, people*

However, because feature vectors are usually very high dimensional, a parametric characterization of their distribution is difficult. In particular, the appropriate number of mixture components (K) is often not known *a prior*, hence necessitating some form of model selection.

In this paper, we explore an alternate, fully Bayesian approach to build an *infinite* mixture model [28], where K is itself a random variable and can potentially be unlimited. Of course, at any point in time, only a finite number of mixture components have data assigned to them. But as more data is seen, more components may be used. Hence, this is a non-parametric model in that the model parameters (e.g., K) grow with the amount (and complexity) of data. This is desirable in our case for visual sense clustering of polysemous images, because different queries inherently have varying levels of ambiguity, leading to different levels of image polysemy. As the complexity (polysemy) of the image data grows, our non-parametric model will also grow to accommodate the greater visual diversity.

We follow recent trends in the vision community to represent images as bags of words. Here, the words refer to both the *textual* words associated with the images, and the *visual* words encoding the visual elements in the images. These two modalities are related since they collectively describe the image content. By incorporating the two modalities in our framework, we exploit their co-occurrence patterns. We adopt the latent topic model [4, 27] to encode these co-occurrence

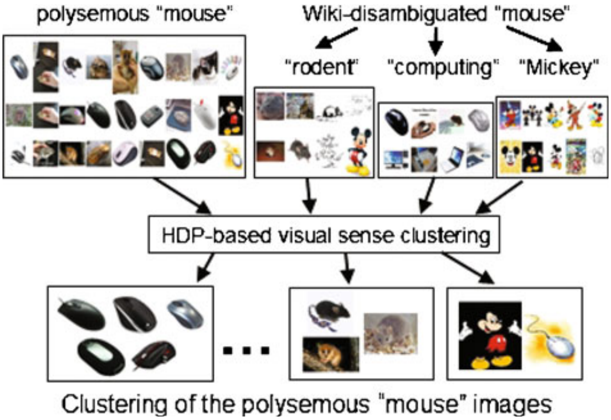
polysemous "mouse"
Wiki-disambiguated "mouse"
"rodent"
"computing"
"Mickey"
HDP-based visual sense clustering
...
Clustering of the polysemous "mouse" images

Fig. 2 The framework of our approach. Given the ambiguous query "mouse", a set of polysemous images is retrieved. Their latent visual senses are elucidated by HDP, helped by a parallel injection of images of suggested semantic senses by Wikipedia

patterns. We define latent topics as multinomial distributions over the words, and make use of the Hierarchical Dirichlet Process (HDP) [34] to learn these topics and derive compact representations of images and clusters.

Our clustering method is completely unsupervised. To constrain the clustering of the images towards their *semantic* senses, we introduce into the clustering process additional images that represent these semantic senses. These *sense-specific* images are obtained by first consulting an online resource (Wikipedia) to solicit possible ways to *expand* the ambiguous query, and then issuing the expanded query to an image search engine. This process of query expansion can be seen as a way to disambiguate query. Together, the polysemous images from the ambiguous query and the sense-specific images from the disambiguated queries are input to the HDP-based clustering step. Collectively, HDP computes latent topics that are shared by all images. Figure 2 shows the framework of our approach.

Our contributions can be summarized as follow:

- We develop a non-parametric visual sense model of web images. We show the efficacy of this model in clustering polysemous images returned from ambiguous search queries. The main advantage of our model is that it grows with the level of polysemy underlying the images.
- Our model exploits the multimodal nature of web images by combining information from their visual content and textual content from the web surround text. Furthermore, to improve semantic clustering, we introduce another related modality of web images, namely their disambiguated senses as suggested by Wikipedia. Using these disambiguated senses, we retrieve sense-specific images and incorporate them into the clustering step.

2 Related works

Image sense elucidation using clustering Our approach of using image clustering as a way to elucidate semantic senses has a few parallels in the literature. By applying spectral clustering on the combined visual features and the text of the embedding web pages, Loeff et al. [21] found that resulting clusters were dominated by images depicting the core (or primary) senses. There was no attempt to elucidate other related senses. Similar to our motivation of computing visual interpretation of a multi-faceted query word, Li et al. [20] proposed correlation analysis techniques, semantic and visual clustering on an image corpus to convey the correct meaning of a concept or word.

In this paper, we ignore iconographic senses (e.g., *apple* on a plate vs *apple* on the tree) and restrict ourselves to only the *core* senses of images (e.g., *apple* the fruit or the company logo). Because the core senses are usually exposed as objects displayed prominently in the images, methods that compute image clusters as distinct object *categories* are relevant to this paper. For example, Philbin et al. [26] computed local features on affine-invariant Hessian regions and applied RANSAC-based spatial verification to detect multiple *instances* of objects in videos and images. Clusters are then formed by querying the objects against a database for similar images. Grauman and Darrell [14] similarly treated unsupervised category learning as an image clustering problem, and proposed an iterative refinement of the primary groupings of images obtained from spectral clustering. They demonstrated superior image classification results by using the images in the resulting clusters to train a visual classifier for that object category. While we share common features (e.g., similar image features) and overlapping end-goals (e.g., to develop better image object models) with the above two papers, the method of clustering used in this paper is very different. Specifically, we use a non-parametric model of clustering that adapts and grows with the complexity of the images. This is especially important as we handle image search queries with different levels of polysemy, resulting in varying levels of complexities in the returned images.

Visual search diversification We aim to cluster images according to their visual senses, and present each of them as a possible interpretation of the original search query. This approach can be considered as a form of *search result diversification*, used by many researchers in the literature as a way to resolve query ambiguity [1, 2, 10]. Ali and Stam first observed the negative impact to movie viewers when given a recommendation list that contains many (near-) duplicates [2]. They coined this negative effect as the "portfolio effect", alluding to the law of diminishing marginal returns commonly known in economics. It explains the decreasing level of enjoyment over a product when it is repeatedly consumed over and over again. Zeigler et al. [46] alleviated this problem by diversifying a result list to reflect the spectrum of user interests. Although their system is detrimental to average accuracy, they show that their method improves user satisfaction. Carbonell and Goldstein [8] proposed a Maximal Marginal Relevance (MMR) ranking function to tradeoff between maximizing relevance while minimizing similarity amongst the retrieved documents. Zhai et al. [45] further extend MMR to a general framework to score documents with probability of relevance and novelty.

It is natural to ask if the above ideas can be extended to the visual domain to diversify visual search. The ImageCLEF 2008/2009 is an international image benchmarking forum that dedicated a specific task on image search diversification [3]. Under this task, given a query topic, images are assigned to clusters depicting subtopics that promotes some predefined *types* of diversity to the original query topic. For example on a query topic asking for "beaches in Brazil", clusters are defined based on *location*; on a topic asking for "animals", clusters are formed based on animal *classes*. Our work in this paper contributes to this line of research by focusing on sense discrimination as a novel diversity type.

There are many existing works that rank images by novelty and relevance. Song et al. [33] used a re-ranking method based on topic richness analysis to enrich topic coverage in retrieval results, while maintaining acceptable retrieval performance. Leuken et al. [18] analyze the visual similarity amongst the images according to their ranking order returned by existing text-based image search engines. Wang et al. [40] evaluated semantic clustering based on a textual analysis of the image search results, while Cai et al. [7] applied multimodal hierarchical clustering to organize web image search results into different semantic groups, using visual cues, textual cues and web link analysis. Wan et al. [39] used a generative model on the keyframes and speech transcripts to enable faceted topic retrieval of news video. The common processing thread in these methods is to add a layer of visual processing to diversify (e.g., to increase novelty in the top-ranked images) the image rank order. While these methods facilitate browsing of image search results from ambiguous queries, the ensuing search results may not necessarily disambiguate the possible sense interpretations of the query.

Latent structures and non-parametric models There has been recent interest in discovering latent visual themes in images using topic models such as the Latent Dirichlet Allocation (LDA) and the hierarchical LDA (hLDA) [4, 5, 12, 27, 32]. LDA mines the feature co-occurrence patterns to uncover the underlying distributions (topics) that best account for the data. Closely related to our work, the latent approach has also been applied to computing image clusters to resolve image polysemy. Given an ambiguous query, Saenko [30] learn an image model to distinguish images of the dominant sense of the query by using the image surround text. Wan et al. [38] extended the model to the visual domain, and learn an image model for every sense suggested by Wikipedia. However, in both work, a nagging issue is that the appropriate number of latent topics is unclear.

By assigning priors with potentially unlimited capacity, the HDP framework allows the design of flexible models to represent complex structures in data. In recent years, development of efficient inference algorithms such as the Markov chain Monte Carlo (MCMC) and variational Bayes have fueled its application to real-world data. Xing et al. [43] used HDP to capture the cross-population structures for multipopulation haplotype inference. Hoffman et al. [15] proposed a similarity measure for songs based on the latent mixture components learnt by HDP. Li et al. [19] adopted HDP as the incremental learning framework for building an image collection. We further explore its utility in resolving polysemy in image search.

3 HDP-based elucidation of image senses

The HDP is an extension of the Dirichlet Process (DP), a type of stochastic process first introduced in the 1960's [13]. But it has recently become an important tool as a prior for infinite mixture models. HDP extends DP in such a way that the dependencies amongst a set of DPs can be specified in a tree structure [34]. HDPs are useful priors for hierarchical mixture models, in which data are organized in groups that share the mixture components. We outline the DP and then describe how we model images and clusters with an HDP.

3.1 Dirichlet process

We begin by considering a Bayesian finite mixture model with J mixture components and each component j has a mixture weight π_j and a parameter vector θ_j. Assume we have N data points denoted x_i, where $1 \leq i \leq N$. Each point is assigned to a mixture component indexed by the indicator z_i. Hence, $p(z_i = j|\pi) = \pi_j$, or $z_i|\pi \sim \text{Multi}(\cdot|\pi)$. The data likelihood is $p(x_i|z_i = j, \theta) = F(x_i|\theta_j)$, where $F(\cdot|\cdot)$ is a pdf parameterized by θ, and $\theta_j \sim H(\cdot)$ for some base distribution H. The mixture weight π has a symmetric Dirichlet prior: $\pi|J, \alpha \sim \text{Dir}(\cdot|\alpha/J)$. From [28], the probability of assigning data point i to component j given all other assignments is:

$$p(z_i = j|z_{-i}, \alpha, J) \propto \frac{n_j^{-i} + \alpha/J}{N - 1 + \alpha} \tag{1}$$

where z_{-i} denotes the assignment of all data points excluding the point i, and n_j^{-i} denotes the number of points assigned to component j excluding point i.

Now we extend our consideration to an infinite mixture model. How can we define a prior for the infinite dimensional parameters? The key is to understand that the infinite dimension weights must *sum to one*. We can construct such a weight distribution by a *stick-breaking* process Stick(α), where α is a concentration parameter. We imagine starting with a stick with unit length and breaking it at a random point. We take the right piece and break it again at a random point. The process is repeated infinitely, producing a set of random weights π that has a *countably* infinite number of dimensions $j = 1 ... \infty$, and whose components all sum to one. H would need to be sampled a countably infinite number of times to generate the component parameter values θ_j.

Now consider the distribution over *all* possible component parameter values θ. This distribution is non-zero at a countably infinite number of values. We denote this distribution by $G(\psi) = \sum_{j=1}^{\infty} \pi_j \delta(\psi - \theta_j)$. Each such G can be seen as a sample from a stochastic process that can be proven to be the DP. In general, a DP is characterized by a scalar parameter α and a base distribution H. A sample from a DP, denoted as $G|\alpha, H \sim \text{DP}(\alpha, H)$, is a distribution that is non-zero over a countably infinite number of values. As we have seen, this is exactly what is required to parameterize an infinite mixture model.

3.2 Hierarchical Dirichlet Process

In the same manner as how DP can act as a prior for infinite mixture models, HDP can also be a prior for the *hierarchical* infinite mixture models. For clarity, we here consider a two-level hierarchy model. But it can be easily generalized to more levels. In HDP, we assume that we have T groups of data, each consisting of N_t data points x_{ti}, $1 \leq t \leq T$ and $1 \leq i \leq N_t$. Each data group is modeled by an infinite mixture model. These models are not independent: the mixtures share component parameters θ and a common DP prior.

The dependencies amongst the infinite mixture models can be again understood using the stick-breaking distribution. Starting at the top level, imagine drawing a sample $G|\alpha_0, H \sim \mathrm{DP}(\alpha_0, H)$. Then we can write as: $G(\psi) = \sum_{j=1}^{\infty} \beta_j^0 \delta(\psi - \theta_j)$ where $\beta^0|\alpha_0 \sim \mathrm{Stick}(\alpha_0)$ are the infinite dimensional mixing weights.

We next form a second DP using G as a base distribution, with scalar parameter α_1, and generate samples from this DP for each of the T mixture models: $G_t|\alpha_1, G \sim \mathrm{DP}(\alpha_1, G)$. Each sample can be written as: $G_t(\psi) = \sum_{j=1}^{\infty} \pi_{tj} \delta(\psi - \theta_j)$ Each G_t inherit the same non-zero points θ_j as G. Hence we have constructed T dependent infinite mixture models. Each model has a separate weight π_t but shared θ.

3.3 Image visual sense model

Following recent trends, we adopt a multinomial bag of words (textual words and visual words) for images. Another motivation for using multinomial distributions is that they are computationally amenable for the latent topic modeling framework [4]. This framework was used in our earlier work in [38] to model the visual senses of images. In this paper, we extend the modeling framework to include HDP as a non-parametric prior. HDP offers the advantage of sample an infinite number of latent topics for each visual sense cluster. This is desirable because different queries inherently have varying levels of ambiguity, leading to different levels of image polysemy. As the complexity of the image data grows, such as when there is greater visual diversity, our HDP model allows us to grow the visual sense model.

Given a collection of polysemous images retrieved from an ambiguous image search query, these images induce a variable number of latent topics proportionate to the number of underlying visual senses (or clusters). Images belonging to the same visual sense (or cluster) are expected to have similar words (textual words and visual words). We now consider each image j as consisting of n_j image feature points $(x_{j1}, \ldots x_{jn_j})$. We assume that these features are exchangeable and to be modeled with a mixture model. While each mixture model has mixing proportions specific to the image, we require that all images share the same set of mixture components. The main idea behind the constraints is to allow statistical strength to be shared amongst images, that also facilitate generalization to new images [34].

We use the HDP as a non-parametric prior to allow component sharing amongst the mixture models. The HDP is a distribution over a set of random probability measures over the parameter space of infinite mixture models. There is one measure G_j for each image j, and a global measure G_0. G_0 is distributed as $\mathrm{DP}(\gamma, H)$, with H the base measure and γ the concentration parameter. Each G_j is conditionally dependent given G_0, with distribution $G_j \sim \mathrm{DP}(\alpha_0, G_0)$.

The goal of learning is to update the parameters in HDP. We adopt the Gibbs' sampling as the learning algorithm [34]. We choose the simpler Chinese Restaurant Franchise (CRF) [34] metaphor to describe the learning process.

Imagine there are an infinite number of Chinese restaurants each with an infinite number of tables (see Fig. 3). All customers sitting on the same table in any restaurant will share the same dish of food. Let x_{ji} be the ith customer in the jth restaurant. In this metaphor, the jth restaurant represents the jth image, and x_{ji} is the ith observed image features. There are n_j such observed features. Recall then that $x_{ji}|\theta_{ji} \sim F(\theta_{ji})$. Hence, for the jth image, there are $\theta_{j1}, \ldots, \theta_{jn_j}$ such parameters. To generate $\theta_{j1}, \ldots, \theta_{jn_j}$, imagine n_j customers (each corresponds to a θ_{ji}) in the

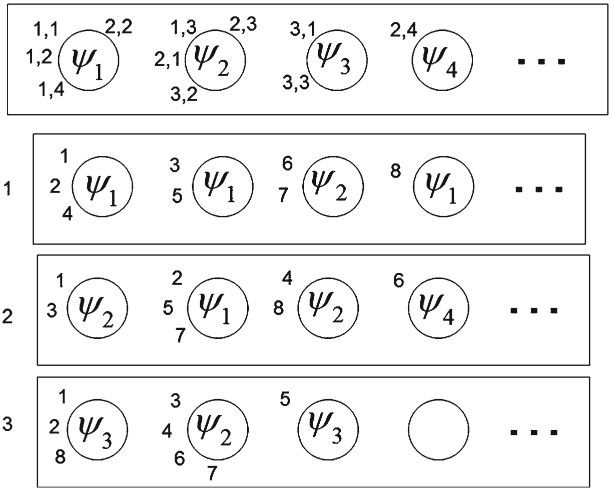

Fig. 3 Chinese Restaurant Franchise (CRF) [34] for three images with eight image feature points. *Top* shows the global restaurant serving dishes shared by all tables in the three local restaurants shown *below*. The generative process in each restaurant is a Chinese Restaurant Process (CRP). For example, in the first local restaurant, the first, third, sixth and eighth customer were the first to sit at an empty table, while the others sat at tables already occupied. For this restaurant, the ninth customer will sit at tables 1, 2, 3 and 4 with probabilities $\frac{3}{8+\alpha_0}$, $\frac{2}{8+\alpha_0}$, $\frac{2}{8+\alpha_0}$, $\frac{1}{8+\alpha_0}$, respectively, or will sit at a new (fifth) table with probability $\frac{\alpha_0}{8+\alpha_0}$. Similarly, in the global restaurant, each customer (j, i) sitting at a table corresponds to table i in the local restaurant j. The table in the global restaurant that (j, i) chooses to sit at, determines the dish that is served at table i in the local restaurant j. In this example, for any new customer entering into a local restaurant j, if he sits down at a new table, then the dish for that new table will be ψ_1, ψ_2, ψ_3 or ψ_4 with probability $\frac{4}{11+\gamma}$, $\frac{4}{11+\gamma}$, $\frac{2}{11+\gamma}$, or $\frac{1}{11+\gamma}$, respectively, or a new dish ψ_5 with probability $\frac{\gamma}{11+\gamma}$

Chinese restaurant. The first customer sits at the first table. A subsequent customer sits at an occupied table with probability proportional to the number of customers already seated there, or at the next unoccupied table with probability proportional to α_0. Suppose customer i sat at table t_{ji}. The above conditional distribution can be written as:

$$t_{ji} \left| t_{j1}, \ldots, t_{ji-1}, \alpha_0 \sim \sum_{t=1}^{T_j} \frac{n_{jt}}{\sum_k n_{jk} + \alpha_0} \delta_t + \frac{\alpha_0}{\sum_k n_{jk} + \alpha_0} \delta_{t^{\text{new}}} \right. \tag{2}$$

where T_j is the current number of tables in restaurant j, and n_{jt} is the number of customers currently sitting at table t. Note that because all customers share the same dish of food, all x_{ji} share the same mixture component indexed by the table t_{ji}. Once all customers have sat down, the seating plan corresponds to a partition of $\theta_{j1}, \ldots, \theta_{jn_j}$. Because this is an exchangeable process [34], the probability of a partition does not depend on the order in which the customers sit down. Now we associate each table t a draw ϕ_{jt} from G_0, and assign $\theta_{ji} = \phi_{jt_{ji}}$.

Performing this process independently for each image j, we have an assignment of each θ_{ji} to a sample $\phi_{jt_{ji}}$ from G_0. Because all ϕ_{jt} are independent draws from

G_0, which is distributed according to DP(γ, H), we may apply the same Chinese restaurant partitioning process to all ϕ_{jt}. This corresponds to partitioning at the next-level hierarchy. Continuing the metaphor, we associate ϕ_{jt} with a customer seated at table k_{jt} at the global level restaurant. However, each k_{jt} is actually a (common) dish that is served to a table at the local restaurant. We can then write the conditional probability as:

$$k_{jt}\left|k_{11}, \ldots, k_{1n_1}, k_{21}, \ldots, k_{jt-1}, \gamma \sim \sum_{k=1}^{K} \frac{m_k}{\sum_{k'} m_{k'} + \gamma}\delta_k + \frac{\gamma}{\sum_{k'} m_{k'} + \gamma}\delta_{k^{\text{new}}}\right. \tag{3}$$

where K is the current number of dishes, and m_k represents the number of local restaurant tables that has ordered dish k. Just as a new table at a local restaurant can be generated from G_0, a new dish at the global level "restaurant" can be generated from H: we draw ψ_k from H and assign $\phi_{jt} = \psi_{k_{jt}}$.

The state space of HDP consists of values of $\boldsymbol{t}$, $\boldsymbol{k}$ and $\boldsymbol{\psi}$. In the global "restaurant", the number of k_{jt} and ψ_k variables is not fixed. We can think of the actual state space as consisting of countably infinite number of k_{jt} and ψ_k. Only finitely many are actually assigned to image feature data and represented explicitly.

Sampling the feature clusters (tables in the local restaurants) Let $f(\cdot|\psi)$ and h be the density functions for $F(\psi)$ and H respectively, n_{jt}^{-i} be the number of tables $t_{ji'}$'s equal to t except t_{ji}, and m_k^{-jt} be the number of dishes $k_{j't'}$'s equal to k except k_{jt}. From (2), we can compute the conditional prior distribution of t_{ji}. Combined with the likelihood of generating image feature x_{ji} given $t_{ji} = t$ (which is simply $f(x_{ji}|\psi_{k_{jt}})$), we obtain the conditional posterior for t_{ji}. Hence, the probability of a new customer x_{ji} sitting at table t is:

$$p(t_{ji} = t|\boldsymbol{t_{-ji}}, \boldsymbol{k}, \boldsymbol{x}, \boldsymbol{\psi}) \propto \begin{cases} \alpha_0 f(x_{ji}|\psi_{k_{jt}}) & \text{if } t = t^{\text{new}} \\ n_{jt}^{-i} f(x_{ji}|\psi_{k_{jt}}) & \text{if } t \text{ is used} \end{cases} \tag{4}$$

Sampling the global latent topics (dishes or tables in the global restaurant) Sampling the k_{jt} variables is similar to sampling the t_{ji} variables described above. We generate a new mixture parameter $\psi_{k^{\text{new}}} \sim H$. Changing k_{jt} changes the component membership of all feature data in table t. The conditional distribution for k_{jt} is:

$$p(k_{jt} = k|\boldsymbol{t}, \boldsymbol{k_{-jt}}, \boldsymbol{\psi}, \boldsymbol{x}) \propto \begin{cases} \gamma \prod_{i:t_{ji}=t} f(x_{ji}|\psi_k) & \text{if } k = k^{\text{new}} \\ m_k^{-t} \prod_{i:t_{ji}=t} f(x_{ji}|\psi_k) & \text{if } k \text{ is used} \end{cases} \tag{5}$$

Sampling $\boldsymbol{\psi}$ Conditioned on the indicator variables $\boldsymbol{k}$ and $\boldsymbol{t}$, ψ_k for each mixture component are mutually independent. The posterior distribution is dependent only on the image feature assigned to component k and is given by:

$$p(\psi_k|\boldsymbol{t}, \boldsymbol{k}, \boldsymbol{\psi_{-k}}, \boldsymbol{x}) \propto h(\psi_k) \prod_{ji:k_{jt_{ji}}=k} f(x_{ji}|\psi_k) \tag{6}$$

3.4 Augmenting the visual sense model with Wiki-sense-disambiguated images

Because of its unsupervised nature, the resulting HDP-based image clusters may not coincide with the semantic meaning of the original query. To constrain the clustering of the images towards their semantic senses, we introduce into the clustering step, images that are exemplars of the various senses of the query. These sense-specific images can also be seen as a kind of "seed" to facilitate clustering [6]. To achieve this, we use Wikipedia to suggest possible ways to expand the ambiguous query. These disambiguated queries are then used to retrieve sense-specific images. Together, the polysemous images from the ambiguous query and the sense-specific images from the disambiguated queries are input to the HDP-based clustering step. Collectively, HDP computes latent topics that are shared by all images.

Wikipedia is a free online encyclopedia, representing the outcome of a continuous collaborative effort of a large number of volunteers. Because of the open and collaborative environment the quality and quantity is well trusted. As a large-scale repository of structured knowledge, Wikipedia is a valuable resource for a diverse array of research activities [24]. One structure of particular interest to this paper is the *disambiguation* page. It gives a detailed list of possible senses (meanings) of ambiguous words by attaching the expression (*disambiguation)* to the name of the ambiguous entity, e.g., *bar_(disambiguation)*, which identifies the disambiguation page of the entity "bar".[2] The advantage of this disambiguation page is that it not only gives the word senses in a structured categorized way, but also links up pages that have further details. All these advantages motivate us to use it for disambiguating keyword based image search. Given an ambiguous query keyword we issue the query to Wikipedia to extract different (senses) meanings of the word automatically.

Because Wikipedia may suggest many superfluous disambiguation, we need a way to assess the informative content of all proposed sense suggestion. There are conceivably many methods to do this, but we choose a simple variant of the method in [35] to evaluate the mutual information content of each suggested disambiguation. Each such disambiguation S comprises of one or more terms $t \in S$ that augment the original query to give it a more specific meaning. We score each candidate disambiguation S by the average *pointwise mutual information* (PMI) of the disambiguated terms $t \in S$ with the original query keyword q, weighted by the relative importance of each term t over the Web corpus [9]. The Mutual Information (MI) of S reflects its "semantic distance" from the ambiguous query q and is computed as follow:

$$\mathrm{MI}(S, q) = \sum_{t \in S} \mathrm{PMI}(t, q|corpus) \times w(t) \tag{7}$$

where $w(t)$ denotes the relative importance of the term t and normalized to sum to 1: $\sum_{t \in S} w(t) = 1$. The PMI between two terms is computed as follow:

$$\mathrm{PMI}(t, q|corpus) = \log \left(\frac{P(t, q|corpus)}{P(t|corpus) \times P(q|corpus)} \right) \tag{8}$$

[2] There is no Wikipedia API to retrieve the disambiguation of a word. However, it is straightforward to write a regexp parser to extract the disambiguation links in Wikipedia pages.

Table 1 Keywords and their Wiki-senses used in our experiments

Keyword	Wikipedia word senses
Bank	Bank finance, Bank building, River Bank, Bank sea floor, Blood bank, Gene bank, Piggy bank
Bar	Bar rod, Bar pole, Dessert Bar, Bar Law, Candy Bar, Barbell
Bass	Bass Drum, Bass guitar, Bass Flute, Bass Fish, Bass Rock, Bass Strait, Bass Instrument, Acoustic Bass Guitar
Mouse	Mouse computing, Mickey Mouse, Mouse Rodent, Stanley Mouse, Mouse anime
Plant	Tree, chemical plant, implant, herb, bush, grass, vines, ferns, mosses, forest
Speaker	Speaker government, loudspeaker, Orator, computer speaker, BBC speaker
Temple	Temple anatomy, hindu temple, mount temple, temple mount, temple Jerusalem
Tiger	Bengal Tiger, Tiger Woods, Tiger Shark, Tiger Snake, Tiger Beer, Tiger Mac OS, Tiger Tank, Tony the Tiger, White Tiger, Detroit Tiger
Watch	Wrist watch, guard, watch tower, wall clock, pocket watch
Window	Window house, computer window, windows operating system, window snyder, window blind

where the probability of term(s) is approximated by maximum likelihood: $P(t, q|corpus) = \frac{\#(t,q)}{\#(corpus)}$ is the fraction of documents in the corpus where both t and q are found, and $P(t|corpus) = \frac{\#(t)}{\#(corpus)}$ is the fraction of documents where the term t is found.

Treating MI in (7) as a form of saliency, we then rank and retrieve the top most salient senses. Because different query keywords have different levels of ambiguity, we cannot simply take the top-N sense suggestions, where N is a fixed number for all query. For example, compared to "mouse", the "tiger" keyword is more ambiguous and hence, can have a longer list of disambiguated senses: say, of size ten. If we mandate N to be ten, then while each of the ten "tiger" senses may be meaningful, some "mouse" senses may be spurious.[3] To resolve this, we define a *single* threshold that is applied to the MI saliency measure for all queries. This threshold is manually set so that we can have a reasonable number of salient senses for each keyword (at least five, and at most ten), and at the same time, the obtained senses are fairly meaningful for all query keywords. Each of the senses is then used to retrieve images from the web using Google Image Search (See Table 1 for examples of automatic disambiguated senses of keywords).

3.5 Presenting final image clusters according to their Wiki-primary senses

Given an ambiguous query, the returned images will comprise of images from various senses. After the HDP parameters are learnt as described in Section 3.3, we have an approximation of the posterior distribution of the latent variables conditioned on the observed image features. This distribution is over the cluster partition assigning the image feature vectors to clusters and a truncated vector π_j defining the mixture proportions of each image j over the finite subset of K mixture components that

[3] For example, besides the more intuitive "Mickey" or "Computing" senses, Wikipedia also suggests "hematoma" to be a "mouse" sense. In boxing usage, a facial "hematoma" is a kind of blood hemorrhage caused by repeated blows to the face.

are actually assigned to image feature points. These mixture components $\psi_{1..K}$ parameterize the latent structures in the image features. Because the HDP machinery enforces statistical sharing of strength, the mixture proportion vectors $\pi_{1..J}$ can compactly express the features for each image $1..J$ in terms of the underlying polysemy in the image collection. That is, we can expect that images corresponding to the same visual sense would have a greater similarity in the π distribution. This is the basis for music similarity described in [15], where the KL divergence of π_i and π_j is used to compute the distance between songs i and j.

Given a query keyword P, we treat the disambiguated senses suggested by Wikipedia as the *primary* senses S_i of P, $i = 1, 2, ..., N_P$, where N_P is the number of Wiki-senses of P. For example, in Fig. 2, the "mouse" keyword has three primary senses, corresponding to the three wiki-disambiguated-senses: "computing", "Mickey" and "rodent". Because these are usually the *semantic* senses of the original query, we propose classifying the polysemous images from the original ambiguous query to one of the N_p primary senses. We define the likelihood of the ith sense S_i given the global latent topic $z = z_j$ as:

$$P(S_i|z = z_j) = \frac{1}{|S_i|} \sum_{a \in S_i} P(a|z = z_j) \tag{9}$$

$$= \frac{1}{|S_i|} \sum_{a \in S_i} \mathrm{KL}(W_a, Z_j) \tag{10}$$

where W_a is the word (concatenated textual and visual) distribution of image a, Z_j is the word distribution of topic z_j, and $\mathrm{KL}(\cdot)$ is the Kullback Leibler divergence between the two. For an image d, the model computes the probability of d belonging to the ith sense S_i as:

$$P(S_i|d) = \sum_{j=1}^{K} P(S_i|z = z_j) P(z = z_j|d) \tag{11}$$

Equation (11) assigns visual sense probabilities to an image according to how similar it is to the sense-specific images. $P(S_i|d)$ provides a way to re-rank the images in the original polysemous order. Images belonging to some sibling senses are given lower probabilities and pushed to the back of the rank list.

4 Experimental results

4.1 Methodology and dataset

In this section, we evaluate how well our visual sense models can distinguish between images depicting the various senses of a given polysemous query keyword. We focus on objects, and define a set of ten polysemous keywords. Two factors motivate our choice of object-based keywords. One, following [21], we focus on the core senses of images, which are typically exemplified as objects displayed predominantly in the images. Two, the three most relevant previous works on sense discrimination [21] and sense model [30, 38] have also focused on object-based keywords. In comparison, our experiments on ten polysemous object keywords provide a more comprehensive evaluation in terms of dataset size and extent of query class ([21] uses three object keywords, while [30] reports on five). For each keyword, we automatically mine the dominant senses from Wikipedia (refer to Section 3.4). As the Web corpus to compute MI saliency, we download and index the English Wikipedia XML dump [42]. Table 1 shows the ten keywords and their respective senses (64 in total).

For each of the ten keywords, we create an image dataset by issuing the keyword as a search query to Google Image Search. We retrieve about 500 images for each keyword. In total, there are 5,013 keyword images. We do the same for the 64 keyword senses, each time issuing to the image search engine with the expanded keywords as search query. We retrieve about 200 images for each keyword sense, totaling 12,336 sense images.

All images were automatically downloaded by following the image URLs on the Google image result index page. For each image, we also retrieve their surround text. To do this, we first remove all HTML and meta-tags, and retain only the non-markup words. Then, centered on the HTML location of the image, we collate all the adjacent 100 words and take them to be the textual content of the image.

Our image labeling procedure is briefly as follow. For each image of a particular keyword, three human labellers were given the list of the word senses, and they were asked to choose only one dominant sense. The dominant sense of an image is the one with majority vote. 27% of images have no majority vote, and they are then labeled as “None” and not used. With three independent opinions, we are assured of some basic coverage of the image senses and the objectivity in their labeling. The extra “None” label is also defined for images that are outside the sense list, or where the object was too small or occluded.

The keyword images and the sense-specific images serve different purposes. For the keyword images, our intent is as follow. Because each keyword is polysemous and can have different meanings, the returned images will contain a mix of images representing the various meanings of the keyword. The efficacy of our visual sense models can be evaluated on how well they can classify each of these images into their respective meaning.

For the sense-specific images, they are used for the following. Firstly, as mentioned in Section 3.4, we constrain our unsupervised clustering of the polysemous images with exemplar images depicting the various senses of the polysemous keyword. Secondly, these sense-specific images are used to create a sense-specific image classifier that becomes a baseline comparison for our visual sense model. For example, by issuing the expanded query “Mouse Computing” to Google Image Search, we can retrieve images of computer-mouse and learn a computer-mouse classifier. Clearly we can compare this classifier to our visual sense model by their classification results on the polysemous “Mouse” images.

4.2 Visual features

We use a dense over-sampling approach to represent images. Local image regions are extracted from three sources: Difference of Gaussian (DoG) interest points [22], Maximally stable extremal regions (MSER) [23], and overlapping rectangular grid. DoG and MSER regions can be viewed as complementary to each other, sampling blob-like regions and high contrast image structures. Similar dense representation has been successfully used in the context of supervised object and scene category recognition [16, 17]. For each region, the SIFT [22] descriptors are then computed using the VLFeat [36] toolkit, and assigned to the nearest visual word from a visual codebook learned on a separate dataset using K-means clustering.

4.3 Baselines

Our goal is to re-rank the polysemous images from the ambiguous keyword query into its various visual senses. Naturally, our first baseline is the existing Google rank of the polysemous images. The choice of this as a baseline seems contrived because by design, Google never intends to rank images into their visual senses. Nonetheless, we retain this baseline for the sake of comparison.

As our second baseline, we use an unsupervised method based on spectral clustering to group images that are iconographically coherent. Spectral methods use eigen-decomposition to compute non-linear clustering of high-dimensional manifold data [25]. In the image domain, they have found successful applications in clustering web images [7, 21]. Spectral clustering works by first constructing a graph Laplacian $L = D^{-1/2}WD^{-1/2}$ of the image data, where W is the pairwise image-similarity matrix, and D is the diagonal matrix such that $D_{ii} = \sum_j W_{ij}$. The first p eigenvectors of L are then taken and arranged as columns in a new matrix Y. The rows of Y are then normalized, and the final cluster assignment is obtained by using the traditional k-means to cluster the rows of Y. The number of k in the final k-means clustering is set to be the same as the number of Wiki-primary senses. Following [25], we encode the affinity matrix W_{ij} between image i and image j as a combination of cosine similarity in text and χ^2 distance between visual words histogram in images: $W_{ij} = \exp(-(1 - \cos^t_{ij}) - (1 - \chi^2_{ij}))$. For brevity, we shall call this baseline method SPEC-CLUST in short.

Because we have sense-specific images by issuing Wiki-disambiguated queries, our third baseline is to bootstrap sense-specific classifiers from these images. We call this method Sense-Specific SVM (SS-SVM in short). While we expect that these images can be more homogeneous as a result of increase query specification, polysemy will nonetheless be a problem in learning the sense-specific SVM (see Fig. 7). In contrast, our approach in this paper resolves these issues by incorporating a latent model of the visual senses of the original polysemous keyword. The key idea is that in these images, there is a rich source of information about the various senses (visual or textual content) of the word. These visual senses capture the salient visual (and textual) characteristics of images associated with the keyword, and offer a more robust model than learning on just the Wiki-sense-specific images.

Our fourth baseline is the visual sense model described in [38]. In that work, a latent image model based on LDA (we shall call it VS-LDA, in short for Visual Sense-LDA) is similarly trained on polysemous images. However, there are two important differences. Firstly, VS-LDA does not make use of the textual content of the images that is provided by the HTML surround text. secondly, for each polysemous keyword, the number of global latent topics (K) is *hard-coded* to be twice the number of Wiki-senses of that keyword. This number is based on the intuition that there are *more* visual topics spanning the polysemous image datasets than that specified by Wikipedia. In contrast, this number is now completely automated by the non-parametric approach described in this paper. We shall call our new model VS-HDP in short.

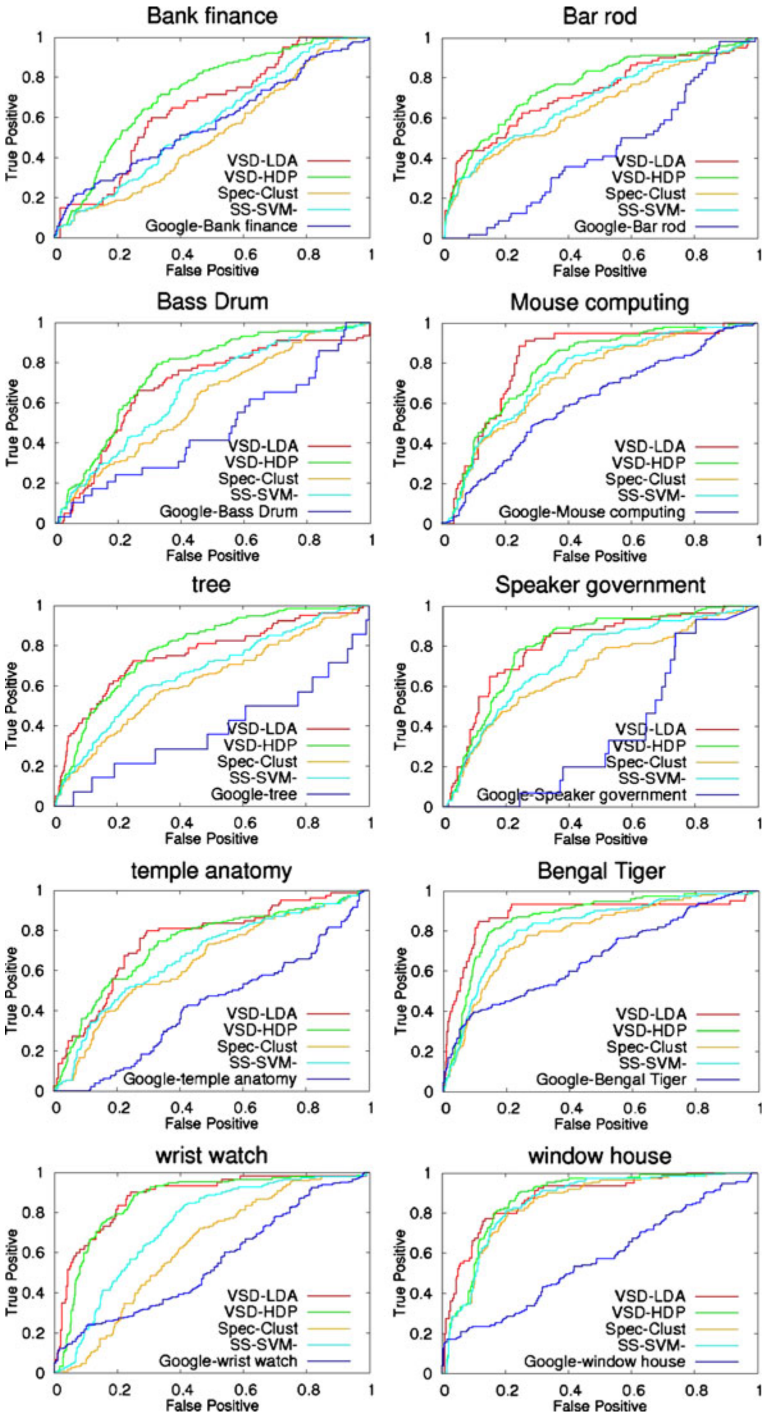

Fig. 4 ROC plots of the first primary sense of the ten polysemous keywords

Table 2 Area Under Curve (AUC) of all senses of each keyword

Keyword	Google rank	SPEC-CLUST	SS-SVM	VS-LDA	VS-HDP
Bank	3.18402	4.8934	5.1769	5.34275	5.50356
Bar	2.48154	3.3478	3.94088	4.20582	4.6842
Bass	3.71719	5.5345	5.47068	5.84983	6.21209
Mouse	1.73422	2.2279	2.41485	2.54619	2.4917
Plant	5.03565	6.8972	7.45235	7.96624	8.06207
Speaker	2.07675	2.4125	2.82301	3.03133	3.04853
Temple	2.1008	2.9134	2.90866	3.01708	2.88802
Tiger	3.87138	6.8976	7.72587	8.11975	8.04058
Watch	2.06972	3.3231	3.7881	4.06134	3.98498
Window	1.93045	3.4665	4.05622	4.19338	4.24679
Total-AUC	28.2017	41.9139	45.7575	48.3337	49.1625

4.4 Evaluation

We now evaluate how well the four algorithms (SPEC-CLUST, SS-SVM, VS-LDA and VS-HDP) can re-rank the polysemous keyword image dataset using a classification approach. For each of the Wiki-primary senses of every query keyword, we train sense-specific image classifiers and apply them to the polysemous images.[4] Images are re-ranked by their classification confidence, where lowly-scored images are moved down to the rank. For example, for each query keyword, a multi-class SS-SVM is trained on the sense-specific images for each sense class of the query keyword. The trained SS-SVM is then used to classify polysemous images into the various senses. For VS-LDA and VS-HDP, we similarly train the two models on the polysemous images. compute $P(S_i|d)$ for each image d using (11), and rank the corresponding images according to the probability of each sense S.

We evaluate the retrieval performance using receiver operating characteristic (ROC) by thresholding $P(S|d)$ for every sense S of a keyword. Due to space constrain, Fig. 4 shows the ROCs for the first Wiki-primary sense of each keyword. The dark-blue lines are the ROCs for the original Google search ranks. The cyan lines are the ROCs using the sense-specific SVMs to re-rank the Google search image order. The orange lines show the ROCs of the SPEC-CLUST baseline, while the red and green lines are the ROCs obtained by VS-LDA and VS-HDP respectively. Table 2 shows the total Area Under Curve (AUC) for all senses of each keyword.

From the results, there are a few notable observations. Firstly, all baseline models, including both VS-LDA and VS-HDP visual sense models, are able to retrieve far more positive class images than the original Google order. While this should not come too much as a surprise, since many authors have also found similar deficiency in Google rank [31], our results are achieved without the need for any training data. Perhaps more surprising is that the re-ranking results of our visual sense model, a

[4] For SPEC-CLUST, our classification-based evaluation framework poses some problem. Because SPEC-CLUST is an unsupervised clustering method, it is unknown which output cluster correspond to the query sense currently being evaluated. We resolve this by manually inspecting the clusters and introspectively labeling the one that best group all images with the query sense. We treat the centroid image of the labeled cluster as the sense prototype, and use the W_{ij} similarity values as classification scores to rank all other images.

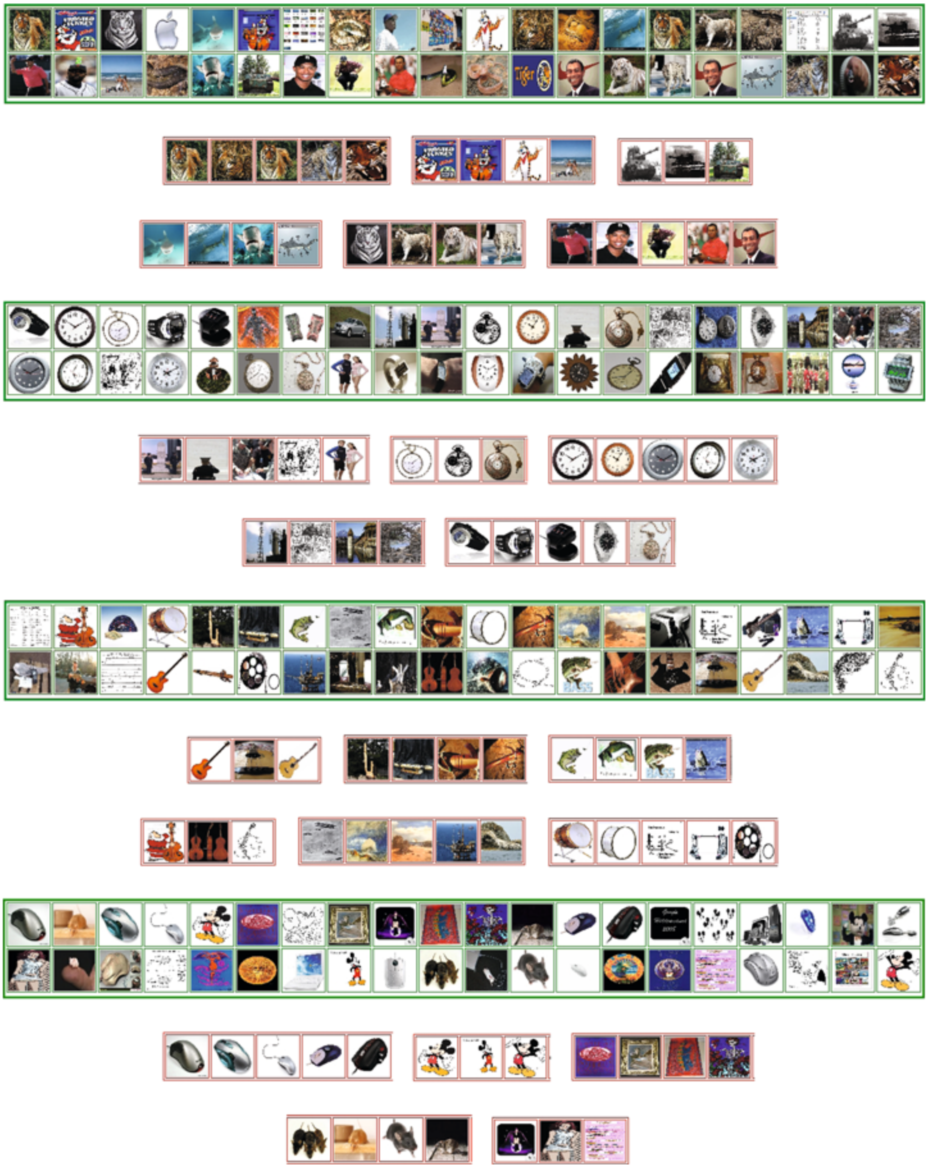

Fig. 5 Example visual sense clusters on four polysemous keywords: "Tiger", "Watch", "Bass" and "Mouse". For each polysemous keyword, we show enclosed in a *green box* two rows of sample polysemous images returned by Google Image Search. Note that because the first Google page of top rank images usually depict the dominant sense of the query keyword, they are fairly homogeneous. Hence, for our illustrative purpose to show sample polysemous images, we randomly take images that Google ranks from 50 onwards. Below the *green box*, we show the automatic clusters of images enclosed in *red boxes*

generative model by design, outperform those of SS-SVM, a discriminative model. We postpone the discussion on this to the next section. Finally, between the two visual sense models, VS-HDP model has produced the better overall performance.

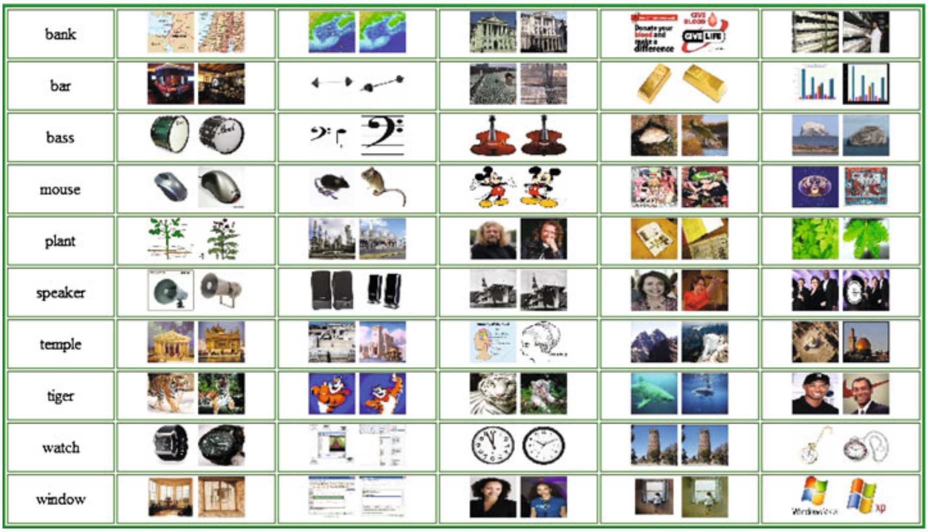

Fig. 6 Visual senses of the ten polysemous keywords. For each keyword, five visual senses are shown. Each visual sense represents a cluster of the images retrieved from the polysemous keyword. From each cluster, we show two images which are visually informative. As can be seen, the primary semantic senses of the polysemous word are captured by the cluster images

This points to the advantage in adapting the sense modeling to the level of polysemy in the images. We show the clustering results on some keywords in Fig. 5. Figure 6 summarizes the first five visual senses for all ten keywords.

4.5 Analysis and discussion

Difficulty for SS-SVM to learn on polysemous images A likely reason for the shortcoming of SS-SVM is that the sense-specific images on which it is trained are also fraught with polysemy (see Fig. 7), making SS-SVM learning difficult. To further illustrate this problem, we show an example of the Wiki-disambiguated query "Mouse-computing" in Fig. 8. In the figure, we compute the average image to visualize the visual polysemy (and diversity) of images. The top figure shows the average image of images retrieved using the expanded query "Mouse-computing".

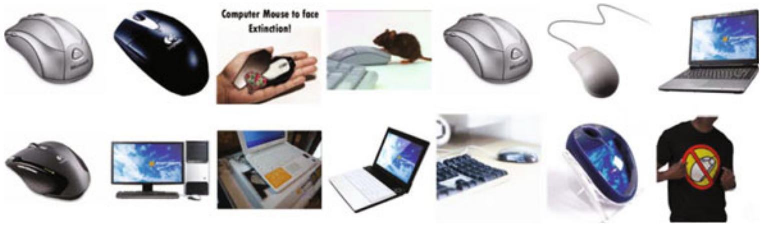

Fig. 7 Google search results on a Wiki-disambiguated query "Mouse computing". While results are more homogeneous than that of "Mouse", polysemy clearly remains an issue

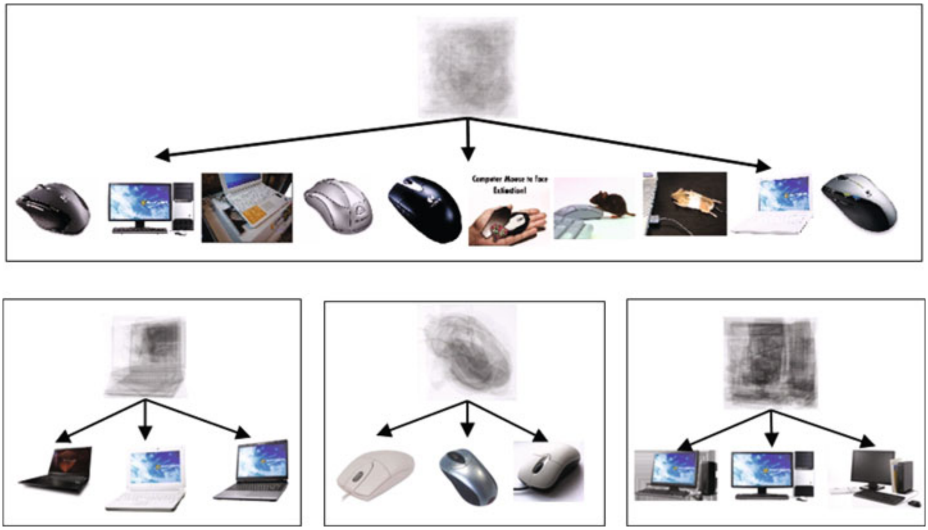

Fig. 8 Using the average image to illustrate the visual polysemy of images. *Top* even for images retrieved using the Wiki-disambiguated query “Mouse computing”, they are very diverse. This will be a challenge to a supervised learning method. *Bottom* average images of three HDP generated clusters on “Mouse computing” images. Note the fine-grained elucidation of “notebook”, “pointer” and “desktop” images

Note that even with the help of this disambiguation so as to retrieve mouse images in the “computing” sense, there is little visual structure in these sense-specific “Mouse-computing” images. In contrast, the images in the three HDP clusters shown in the bottom figure can be seen to exhibit more visual structures. Using HDP, further fine-grained image senses are elucidated, wherein “Mouse-computing” images are now clustered into the “notebook”, “pointer-device” and “desktop” senses.

Effect of K on AUC results We take a closer look at the two most competitive models, the VS-LDA and VS-HDP. While the overall performance difference of two models over the ten keywords are statistically insignificant, the main advantage of the VS-HDP model is that it circumvents the need for a prior choice of the number of mixture components K. In contrast, the VS-LDA model needs to define K a-priori, and the optimal value of K is necessarily determined via repeated trials and cross-validation. We show this in Fig. 9 on two query words (“Bass” and “Speaker”), but results are similar for the other query words. We plot the comparative performance of the parametric models against the non-parametric VS-HDP across varying K. Apart from VS-LDA, we also use a truncated version of the VS-HDP model[5] as

[5] Note that in HDP, even though the CRP prior allows an infinite multinomial distribution over the mixture components, each image nevertheless learns a posterior distribution $\hat{\pi}$ over only a *finite* subset of cluster partition that are actually assigned to image feature vectors. Following [15], we use a truncated DP approach, and truncate β at $K+1$, so that $\beta_k = 0$ for all $k > K+1$. Hence, $\hat{\pi}_j \sim \mathrm{DP}(\alpha^{\pi}, \beta_{1,..,\infty})$ becomes $\hat{\pi}_j \sim \mathrm{Dirichlet}(\alpha^{\pi}, \beta_{1,...,K+1})$. Specifically, for each image j, we let $\pi_{j,1..K} = \hat{\pi}_{j,1..K}$, and $\pi_{j,K+1} = 1 - \sum_{k=1}^{K} \hat{\pi}_{j,k}$, where $\pi_{j,K+1}$ denotes the probability of sampling from a mixture component that has not been used to explain any image feature vector.

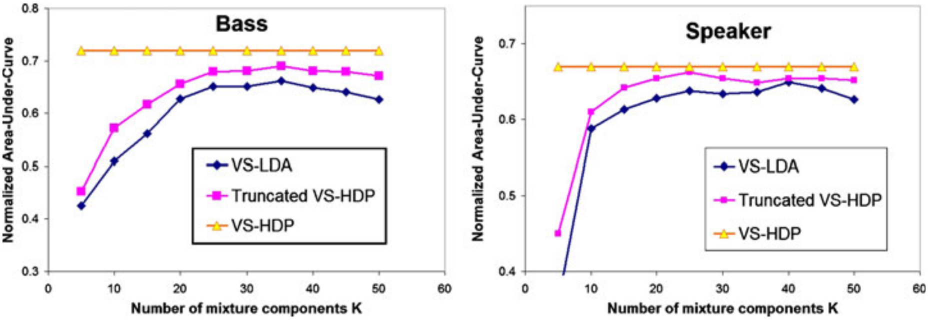

Fig. 9 Comparative AUC performance across various values of the number of mixture components

another parametric model. Note the different values of K for which their normalized AUC is nearest to the best obtained from VS-HDP: 35 for "Bass", and 25 for "Speaker". The higher K for "Bass" reflects its greater ambiguity than "Speaker" ("Bass" has eight Wiki-senses, while "Speaker" has five). For VS-LDA, its AUC results are most sensitive to the value of K. It peaks around the empirical optimal point before deteriorating.

Effect of injecting sense-disambiguated images The appeal of unsupervised clustering methods is that they have no need for manual labels. However, their resulting clusters are generally of lower quality than those generated with supervision. On the other hand, it is often possible to improve clustering results by including some form of limited supervision. For example, pairwise constraints are often used to indicate that two particular data points should belong to same (or different) clusters. In this paper, the use of sense-disambiguated images (suggested by Wikipedia disambiguation links) can be seen as an injection of supervised labels onto the unsupervised HDP clustering framework. The inclusion of these sense-specific images constrain the clustering so that fine-grained sense clusters can be obtained (see Fig. 8). On first thought, it would seem that these sense-specific images already encode the semantic senses that we want. However, this is not the case, since even with these expanded queries, returned images are not homogeneous. The important thing to note is that it is not just the inclusion of *more* data that matters (because much of these data are also noisy), it is the judicious combination of the adaptive strength-sharing

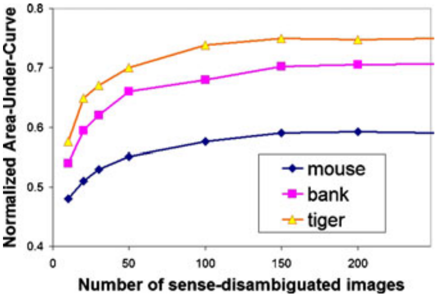

Fig. 10 Including more sense-specific images improve AUC performance of VS-HDP

machinery of HDP, that results in the fine-grained sense clustering. Figure 10 charts the performance improvement as more sense-specific images are used to constrain VS-HDP clustering. AUC results rapidly improve early on but plateau off when hundreds of images are used. We note that this is the range where the return Google images start to be of lower precision, i.e., noisier. In our dataset, we crawled 200 sense-specific images for each keyword sense (see Section 4.1).

5 Conclusion

We develop a method that learns the visual sense model of images. By using it to classify polysemous images into their semantic categories, we are able to extend Rijsbergen's cluster hypothesis to the image domain.

We relate the lexical ambiguity of a query keyword to the corresponding polysemy in its return images. The inherent *word-based* senses of the query is reflected in the *image-based* senses: the more ambiguous the query word, the greater the visual polysemy (and diversity) in the retrieved images. By clustering these polysemous images into their image senses, we present the clusters as embodying the possible visual interpretations of the query keywords. We coin the collective images in these clusters as the *visual senses* of the query keywords, and show how these sense clusters can be used as a way to diversify and disambiguate image search queries.

Our method is completely unsupervised. It capitalizes on the large amount of unlabeled images available through keyword image search to learn a generative model of sense. We extend the notion of image multimodality to include its dictionary senses. We not only exploit the conventional textual and visual information in web images, we also incorporate a list of suggested disambiguated *senses* from Wikipedia. These sense suggestions act as a form of query expansion to solicit further sense-specific images, and inject limited supervised labels into our unsupervised clustering framework. Collectively, the three modalities guide the development of robust sense models for images.

Compared to conventional generative models with fixed prior, we inject flexibility in our visual sense model by using the Hierarchical Dirichlet Process (HDP) as a non-parametric prior, with potentially unlimited capacity. The HDP framework allows our model to be adaptive to the level of polysemy of the image data, which is in turn related to level of polysemy in the query.

We use our visual sense models to classify polysemous images into their sense clusters, according to their sense probability. On a large dataset of images consisting of search results from ten polysemous keywords, our visual sense models improve on both the baseline (Google) search engine, a state-of-art spectral-clustering method, and bootstrapping SVMs trained on the sense-specific images.

Acknowledgements The authors would like to thank Sujoy Roy, Yap Ghim Eng, Sim Tze Jan, Sim Khe Chai and Wang Yue for valuable discussions, and three student helpers for their labeling effort.

References

1. Agrawal R, Gollapudi S, Halverson A, Ieong S (2009) Diversifying search results. In: Proc of the second ACM international conference on web search and data mining, pp 5–14

2. Ali K, Stam V (2004) TiVo: making show recommendations using a distributed collaborative filtering architecture. In: Proc ACM international conference on knowledge discovery and data mining, pp 394–401
3. Arni T, Clough P, Sanderson M, Grubinger M (2008) Overview of the ImageCLEFphoto 2008 photographic retrieval task. In: Working notes of the 2008 CLEF workshop
4. Blei D, Ng A, Jordan M (2003) Latent dirichlet allocation. J Mach Learn Res 3:993–1022
5. Blei D, Griffiths T, Jordan M (2010) The nested chinese restaurant process and Bayesian nonparametric inference of topic hierarchies. J ACM 57(2):1–30
6. Bradley P, Fayyad U (1998) Refining initial points for k-means clustering. In: Proc international conference on machine learning, pp 91–99
7. Cai D, He X, Li Z, Ma W, Wen J (2004) Hierarchical clustering of WWW image search results using visual, textual and link information. In: Proc multimedia, pp 952–959
8. Carbonell J, Goldstein J (1998) The use of MMR, diversity-based reranking for reordering documents and producing summaries. In: Proc ACM SIGIR conference on research and development in information retrieval, pp 335–336
9. Cilibrasi R, Vitanyi P (2007) The google similarity distance. IEEE Trans Knowl Data Eng 19(3):370–383
10. Clarke C, Kolla M, Cormack G, Vechtomova O, Ashkan A, Buttcher S, MacKinnon I (2008) Novelty and diversity in information retrieval evaluation. In: Proc ACM SIGIR conference on research and development in information retrieval, pp 659–666
11. Cutting D, Karger D, Pedersen J, Tukey J (1992) Scatter/gather: a cluster-based approach to browsing large document collections. In: Proc ACM SIGIR conference on research and development in information retrieval
12. Fergus R, Li F, Perona P, Zisserman A (2005) Learning object categories from googles image search. In: Proc international conference on computer vision
13. Ferguson T (1973) A Bayesian analysis of some nonparametric problems. Ann Stat 1:209–230
14. Grauman K, Darrell T (2006) Unsupervised learning of categories from sets of partially matching image features. In: Proc computer vision and pattern recognition
15. Hoffman M, Blei D, Cook P (2008) Content-based musical similarity computation using the hierarchical dirichlet process. In: Proc international conference on music information retrieval
16. Jurie F, Triggs B (2005) Creating efficient codebooks for visual recognition. In: Proc international conference on computer vision, vol 1, pp 604–610
17. Lazebnik S, Schmid C, Ponce J (2006) Beyond bags of features: spatial pyramid matching for recognizing natural scene categories. In: Proc conference on computer vision and pattern recognition, pp 2169–2178
18. Leuken V, Reinier H, Garcia L, Olivares X, Roelof V (2009) Visual diversification of image search results. In: Proc of the 18th international conference on world wide web, pp 341–350
19. Li L, Wang G, Li F (2007) Optimol: automatic object picture collection via incremental model learning. In: Proc computer vision and pattern recognition
20. Li H, Tang J, Li G, Chua T (2008) Word2image: towards visual interpreting of words. In: Proc ACM international conference on multimedia, pp 813–816
21. Loeff N, Alm C, Forsyth D (2006) Discriminating image senses by clustering with multimodal features. In: Proc COLING/ACL, pp 547–554
22. Lowe D (2004) Distinctive image features from scale-invariant keypoints. J Comput Vis 60(2): 91–110
23. Matas J, Chum O, Urba M, Pajdla T (2002) Robust wide baseline extremal regions. In: Proc British machine vision conference, pp 384–396
24. Mihalcea R (2007) Using Wikipedia for automatic word sense disambiguation. In: Proc the annual conference of the North American Chapter of the Association for Computational Linguistics
25. Ng A, Jordan M, Weiss Y (2001) On spectral clustering: analysis and an algorithm. In: Advances in neural information processing systems, vol 14, pp 849–856
26. Philbin J, Chum O, Isard M, Sivic J, Zisserman A (2007) Object retrieval with large vocabularies and fast spatial matching. In: Proc computer vision and pattern recognition
27. Quelhas P, Monay F, Odobez J, Gatica-Perez D, Tuytelaars T, Gool LV (2005) Modeling scenes with local descriptors and latent aspects. In: Proc international conference on computer vision
28. Rasmussen C (2000) The infinite gaussian mixture model. In: Neural information processing systems
29. Rijsbergen C (1979) Information retrieval. University of Glasgow

30. Saenko K, Darrell T (2008) Unsupervised learning of visual sense models for polysemous words. In: Proc neural information processing systems
31. Schroff F, Criminisi A, Zisserman A (2007) Harvesting image databases from the web. In: Proc international conference on computer vision
32. Sivic J, Russell B, Zisserman A, Freeman W, Efros A (2008) Unsupervised discovery of visual object class hierarchies. In: Proc computer vision and pattern recognition
33. Song K, Tian Y, Gao W, Huang T (2006) Diversifying the image retrieval results. In: Proc multimedia, pp 707–710
34. Teh Y, Jordon M, Beal M, Blei D (2007) Hierarchical dirichlet processes. J Am Stat Assoc 101(476):1556–1581
35. Turney P (2002) Thumbs up or thumbs down? semantic orientation applied to unsupervised classification of reviews. In: Proc association of computational linguistics, pp 417–424
36. Vedaldi A, Fulkerson B (2008) VLFeat: an open and portable library of computer vision algorithms. http://www.vlfeat.org/. Accessed Sep 2009
37. Vivisimo (2009) Vivisimo web clustering. http://vivisimo.com/. Accessed Jan 2010
38. Wan K, Tan A, Lim J, Chia L (2009) A latent model for visual disambiguation of keyword-based image search. In: Proc british machine vision conference
39. Wan K, Tan A, Lim J, Chia L (2010) Faceted topic retrieval of news video using joint topic modeling of visual features and speech transcripts. In: Proc international conference on multimedia and expo
40. Wang S, Jing F, He J, Du Q, Zhang L (2007) Igroup: presenting web image search results in semantic clusters. In: Proc of the SIGCHI conference on Human factors in computing systems, pp 587–596
41. Wei X, Croft W (2006) LDA-based document models for ad-hoc retrieval. In: Proc ACM SIGIR conference on research and development in information retrieval, pp 178–185
42. Wikipedia (2010) English dumps in SQL and XML. http://download.wikimedia.org/enwiki/20100116/. Accessed Feb 2010
43. Xing E, Sohn K, Jordan M, Teh Y (2006) Bayesian multi-population haplotype inference via a hierarchical dirichlet process mixture. In: Proc international conference on machine learning
44. Zeng H, He Q, Chen Z, Ma W, Ma J (2004) Learning to cluster web search results. In: Proc ACM SIGIR conference on research and development in information retrieval
45. Zhai C, Cohen W, Lafferty J (2003) Beyond independent relevance: methods and evaluation metrics for subtopic retrieval. In: Proc ACM SIGIR conference on research and development in information retrieval, pp 10–17
46. Ziegler C, Mcnee S, Konstan J, Lausen G (2005) Improving recommendation lists through topic diversification. In: Proc international conference on world wide web, pp 22–32

Kong-Wah Wan received his B.Sc. (Hons) and M.Sc. degrees in Computer Science from the National University of Singapore. He is currently a Research Manager with the Institute for Infocomm Research (I2R), and a lead investigator of a few projects examining image and video retrieval. His work has been used in several large scale deployments in Singapore, including the Port of Singapore Authority (PSA Corp) and the Immigration and Checkpoint Authority (ICA) of Singapore. A few of his recent projects in video content analysis are also exhibited at the STARHome, a technology showcase of the Agency for Science, Technology and Research in Singapore. He has five patents (awarded and pending) and is a recipient of the Tan Kah Kee Young Inventor Award (Silver) in 2004.

Ah-Hwee Tan is an Associate Professor and the Division Head of Information Systems at the School of Computer Engineering (SCE), Nanyang Technological University. He has been a faculty member of SCE since 2003 and was the founding Director of Emerging Research Laboratory, a research center for incubating new interdisciplinary research initiatives. Prior to joining NTU, he was a Research Manager at the A*STAR Institute for Infocomm Research (I2R), spearheading the Text Mining and Intelligent Agents research programmes. His current research interests include intelligent agents, cognitive and neural systems, machine learning, knowledge discovery and text mining.

Prof. Tan received a Ph.D. in Cognitive and Neural Systems from Boston University, a Bachelor of Science (First Class Honors) (1989) and a Master of Science (1991) in Computer and Information Science from the National University of Singapore. He is a recipient of Lim Soo Peng Book Prize, Asia Life Gold Medal, Cambridge Scholarship, Tan Kah Kee Young Inventor Award (Silver), NUS Overseas Graduates Scholarship, KRDL High Achiever Award, Optimal 2003 Gold Award, and Tan Chin Tuan Fellowship. He is an Editorial Board Member of Applied Intelligence published by Springer-Verlag, a Member of ACM, and a Senior Member of IEEE.

Joo-Hwee Lim received his B.Sc. (Hons I) and M.Sc. (by research) degrees in Computer Science from the National University of Singapore and his Ph.D. degree in Computer Science & Engineering from the University of New South Wales. He has joined Institute for Infocomm Research (I2R) and its predecessors since October 1990. He has conducted research in connectionist expert systems, neural-fuzzy systems, handwriting recognition, multi-agent systems, and content-based retrieval. He was a key researcher in two international research collaborations, namely the Real World Computing Partnership funded by METI, Japan and the Digital Image/Video Album project with CNRS, France and School of Computing, National University of Singapore. He also contributed technical solutions to a few industrial projects involving pattern-based diagnostic tools for aircraft and battleship navigation systems and knowledge-based post-processing for automatic fax/form recognition. He has nine patents (awarded and pending) and published more than one hundred and twenty refereed international journal and conference papers in his research areas. He is currently the Department Head of the Computer Vision & Image Understanding Department, with staff strength of fifty research scientists and engineers, at I2R, Singapore. He is also the co-Director of IPAL (Image Perception, Access and Language), a French-Singapore Joint Lab (UMI 2955, January 2007–December 2010). He is bestowed the title of 'Chevallet dans l'ordre des Palmes Academiques' by the French Government in 2008.

Liang-Tien Chia received his B.Sc. degree in Electrical and Electronics Engineering and Ph.D. from Loughborough University (of Technology) in 1990 and 1994, respectively. During this period, he was awarded the University Traveling Prize for his academic achievements and he was a recipient of the Overseas Research Scholarship Award.

He is currently an Associate Professor in the School of Computer Engineering, Nanyang Technological University, Singapore. He was the Director of the Centre for Multimedia and Network Communications from 2002 to 2007 and is currently Head of the Division of Computer Communications.

Liang-Tien Chia's research interests can be broadly categorized into two main areas, Internet related research with emphasis on the Semantic Web and Multimedia Understanding for Information Management through media analysis, annotation and adaptation. Related topics include multimedia storage & retrieval, multimedia processing, multimodal data fusion and multimedia adaptation/transmission.

He is involved in a number of funded research projects and he serves as a member on numerous conference program committees and reviews for some international journals and he has published over 100 referred international conference and journal papers. He was awarded the A*Star Overseas Attachment Programme and spent one month in Microsoft Research Asia.

Liang-Tien Chia is a council member of the Infocomm Technology Standards Committee and serves as the technical chairman of ITSC Plugfest 2006. As a Singaporean, he is an active National Serviceman serving as a Signal Officer in the Singapore Armed Forces.