Social Computing for Collaborative Image Understanding

Jianping Fan\textsuperscript{1}, William Ribarsky\textsuperscript{1}, Ramesh Jain\textsuperscript{2}

\textsuperscript{1}Department of Computer Science, University of North Carolina, Charlotte, NC 28223, USA.
\textsuperscript{2}Bren School of Information and Computer Sciences, University of California, Irvine, USA

ABSTRACT

With the advance of the Internet and the increasing accessibility of computing resources, humans and computer systems are now brought together in powerful new ways. In this paper, we propose a human-centered computing framework to harness the essential characteristics of both humans and computers for achieving collaborative image understanding (i.e., training large numbers of inter-related classifiers collaboratively for automatic object and concept detection from images), where groups of volunteers may collaborate on: (a) giving their timely guidances for supporting collaborative classifier training; (b) using their personal computing resources such as PCs for training large numbers of inter-related classifiers collaboratively; and (c) assessing the correctness of learning results (classifiers and their decision boundaries) and the effectiveness of hypotheses for classifier training.

Keywords: Social computing, collaborative image understanding, human-centered computing, classifier training and assessment.

1. Introduction

Enabling humans to efficiently transfer their knowledge and skills to computer systems has inspired decades of research, and computer systems with humankind intelligence at certain levels have become pervasive within socio-physical contexts. For examples, government agencies and large companies have used computer systems to replace human telephone operators in services and automated transaction machines (ATM) have already replaced many functions of bank tellers. Even developing computer systems with humankind intelligence has obtained some surprising successes over the decades, it still suffers from many failings and it is still one of the greatest challenges in computer science. For
the task of automatic image understanding (automatic object and concept detection from images), humans may significantly outperform computer systems [1-7]. On the other hand, computer systems have strong computational power and large storage memory and they can be used to perform the tasks that humans are inherently not good at (for example, extracting high-dimensional visual features from large-scale images) [1-3]. Thus it is very attractive to develop human-centered multimedia computing platforms that are able to harness the essential characteristics of both humans and computer systems.

We are witnessing the creation of richly interconnected worlds where humans and computer systems together demonstrate new forms of collaboration and emergent intelligence, which were not previously achievable by humans or computer systems along [13-14]. For examples, large numbers of volunteers collaboratively write encyclopedia articles of unprecedented scope and scale [11], create open softwares [12], and tag images/videos [8-10]. Human-centered computing has become a central theme across many research fields [4-7], but most existing systems have not sufficiently harnessed the essential characteristics of both humans and computer systems. On the other hand, most published techniques for automatic image understanding and annotation [1-3] are targeted on automatic machine learning on a single PC, but they have not sufficiently exploited both the computational power of millions of unused computers worldwide and the specific abilities of massive numbers of human participants for training large numbers of inter-related classifiers collaboratively. Thus it is very attractive to develop human-centered multimedia computing platforms that can gracefully scale from a single user to a collaborative environment and from a single PC to large numbers of PCs on the Internet.

People may have strong collaboration motivations on supporting collaborative image understanding (i.e., training large numbers of inter-related classifiers collaboratively for automatic object and concept detection from images) because of the following reasons: (1) they may deal with the same challenging issue of automatic image understanding (automatic object and concept detection from images) and close collaboration can allow them to build a social group to share their common interest and keep track of the advances of collaborative image understanding; (2) they may not be able to solve the challenging issue of automatic image understanding individually because the number of object classes and image concepts (that are needed to be detected) could be very large; and (3) they may want to collaborate and compete each other on their timely achievements. By bringing humans and computer systems together in powerful new ways, social computing may provide multiple unique and innovative ways to achieve collaborative image understanding.
In this paper, we propose a human-centered computing framework to harness the essential characteristics of both humans and computers and leverage large-scale collaboratively-tagged images for achieving collaborative image understanding. To harness humans’ strong capabilities on pattern recognition and hypothesis making for collaborative image understanding, we propose a human-centered computing framework to leverage the specific abilities of massive numbers of human participants for training large numbers of inter-related classifiers collaboratively, where groups of volunteers can collaborate on: (a) giving their timely guidances for supporting collaborative classifier training; (b) using their personal computing resources such as PCs for training large numbers of inter-related classifiers collaboratively; and (c) assessing the correctness of learning results (classifiers and their decision boundaries) and the effectiveness of hypotheses for classifier training. To harness the computational power of millions of unused computers worldwide for collaborative image understanding, we propose a collaborative computing framework to achieve more effective analysis of large-scale image collections and gain deep insights quickly. By making humans’ guidances and computers’ achievements to be transparent, our human-centered computing framework can allow humans to communicate and collaborate more effectively and enable computers to leverage human guidances for improving classifier training, so that we can solve the challenging issue of automatic image understanding collaboratively.

2. Concept Network for Organizing Training Tasks and Volunteer Collaborations

Collaborative image tagging [8-10] has become very popular for obtaining large-scale weakly-labeled images by leveraging the collaborative efforts of a large population of Internet users. Large-scale collaboratively-tagged images can provide multiple advantages: (1) they can characterize diverse visual properties of object classes and image concepts more sufficiently; (2) they can be obtained easily by leveraging the collaborative efforts of large numbers of Internet users, our fundamental belief is that a large group of Internet users with diverse backgrounds can do better job than a small team of professionals as illustrated by wikipedia [11]; (3) both their tags and their visual properties are diverse, thus they can give a real-world point of departure for achieving automatic object and concept detection from images (i.e., automatic image understanding). We have collected large-scale collaboratively-tagged images from Flickr and we have also developed multiple practical techniques to make such collaboratively-tagged images to be useful for classifier training [20-23].

In a collaborative image tagging space, people may use large vocabulary of text terms for image tagging, thus not all the social tags are meaningful for object and concept interpretation, e.g., some
social tags may not have exact correspondences with the real-world object classes and image concepts. To determine the social tags for object and concept interpretation, we first partition the social tags into two categories: *tags of interest* and *tags of non-interest*. The tags of interest are used to construct concept network because the corresponding real-world object classes and image concepts are significant and popular in large-scale collaboratively-tagged images. On the other hand, the tags of non-interest are not used for concept network construction because the corresponding object classes and image concepts are less popular. We do not claim that the less popular object classes and image concepts are not important, but it could be hard to collect enough training images for learning their classifiers reliably.

For two given tags of interest (i.e., two popular object classes or image concepts) $C_i$ and $C_j$ and their image instances, their inter-concept visual similarity context $\gamma(C_i, C_j)$ is determined by cumulating the pairwise visual similarity contexts between their image instances [23]. In this work, multiple criteria (both flat and hierarchical contexts) are leveraged to achieve more precise characterization of inter-concept semantic similarity contexts in a collaborative image tagging space. For two given object classes or image concepts $C_i$ and $C_j$, their inter-concept semantic similarity context $\phi(C_i, C_j)$ consists of two components: (1) *flat inter-concept semantic similarity context* because of their co-occurrences in large-scale collaboratively-tagged images (e.g., higher co-occurrence probability corresponds to stronger inter-concept semantic similarity context) [20-21]; and (2) *hierarchical*
inter-concept semantic similarity context because of their inherent correlation defined by WordNet (e.g., closer on WordNet [19] corresponds to stronger inter-concept semantic similarity context).

For two given object classes or image concepts $C_i$ and $C_j$, a cross-modal similarity alignment framework is developed to determine their cross-modal inter-concept similarity context $\varphi(C_i, C_j)$ by: (a) treating the inter-concept semantic similarity context $\phi(C_i, C_j)$ and the inter-concept visual similarity context $\gamma(C_i, C_j)$ as two different views of the cross-modal inter-concept similarity context $\varphi(C_i, C_j)$; (b) projecting both the inter-concept semantic similarity context $\phi(C_i, C_j)$ and the inter-concept visual similarity context $\gamma(C_i, C_j)$ onto a comparable space and finding the optimal projection directions to make their correlations to be mutually maximized; and (c) aligning the inter-concept semantic similarity context $\phi(C_i, C_j)$ with the inter-concept visual similarity context $\gamma(C_i, C_j)$ on the optimal projection direction to obtain the cross-modal inter-concept similarity context $\varphi(C_i, C_j)$.

Aligning the inter-concept semantic similarity contexts with the inter-concept visual similarity contexts is still an unexplored issue, thus human perceptual factors may play an important role in the design of such the cross-modal similarity alignment frameworks. Based on this observation, a hyperbolic concept network visualization algorithm [18, 20] is incorporated to enable interactive concept network exploration as shown in Fig. 1, so that users can assess the correctness of such cross-modal inter-concept similarity contexts interactively. The timely assessment feedbacks from users can be leveraged for defining more suitable forms to achieve more accurate alignment between the inter-concept semantic similarity contexts and the inter-concept visual similarity contexts. Thus our algorithm can leverage both the strong computational power of computers and the strong pattern recognition capabilities of humans for achieving more precise construction of the concept network.
After the concept network is constructed, it is then used to identify inter-related learning tasks automatically. As shown in Fig. 2, the first-order nearest neighbors on our concept network are used to determine the inter-related object classes and image concepts. It is worth noting that the classifiers for such inter-related object classes and image concepts are strongly inter-related and they should be trained jointly rather than independently. To achieve more effective training of such inter-related classifiers, a pairwise approach is used for SVM classifier training and combination, e.g., each pairwise SVM classifier focuses on distinguishing one particular pair of such inter-related object classes and image concepts. The task for training such pairwise SVM classifier can be accomplished more effectively because the hypothesis space for one particular object/concept pair may have smaller variance and less uncertainty. Thus identifying the inter-related object classes and image concepts and learning their pairwise SVM classifiers jointly can bring more powerful inference schemes to enhance the discrimination power of inter-related classifiers significantly. By incorporating our concept network for training task organization, our pairwise approach for SVM classifier training and combination can provide a good environment to organize groups of volunteers and their collaborations and communications for collaborative classifier training.

3. Collaborative Classifier Training

The object classes and image concepts are dependent and may share some common visual properties (i.e., inter-concept visual similarity). Isolating such inter-related object classes and image concepts and learning their classifiers independently may seriously affect their discrimination power. On the other hand, the learning complexity for some object classes and image concepts could be very high, we may need large numbers of training images to achieve reliable training of their classifiers. As a result, it is very hard if not impossible to use a single PC to store large numbers of training images and train multiple inter-related classifiers jointly. Another drawback for most existing machine learning techniques is that they have not sufficiently leveraged human guidances for improving classifier training [1-3]. Without involving human beings to make more suitable hypotheses and provide more precise assessments, it is very hard for most existing machine learning techniques to achieve reliable classifier training. Thus it is very attractive to develop new human-centered computing frameworks that are able to harness groups of volunteers and their unused PCs and their guidances to enable collaborative classifier training.
In this work, a concept network is generated for more than 1000 most popular object classes and image concepts, where each object class and image concept contains more than 5000 image instances for achieving more reliable classifier training. When the object classes and image concepts are inter-related on the concept network, the tasks for training their classifiers are strongly dependent, thus their image instances should be integrated to train their inter-related classifiers jointly rather than independently. As a result, it is very hard if not impossible by using a single PC to store all the relevant image instances simultaneously in the memory and train multiple inter-related classifiers jointly.

To achieve collaborative training of more than 1000 inter-related classifiers, a collaborative computing framework is developed and it consists of the following key components as shown in Fig. 3:

(a) The tasks for training more than 1000 inter-related classifiers are decomposed into groups of inter-related learning tasks, where groups of volunteers are involved to jointly learn the inter-related classifiers for the inter-related object classes and image concepts on the concept network. For a given object class or image concept, we first consider its first-order nearest neighbors on the concept network to determine the inter-related learning tasks and some examples are shown in Fig. 2.

(b) For each training group, all its inter-related learning tasks are distributed among a group of volunteers and the relevant image instances are also distributed among their unused PCs. The concept network is also used to organize the collaborations and communications. The inter-related classifiers, which are trained collaboratively by groups of volunteers, can be integrated for supporting automatic image understanding. For a given object class or image concept $C_j$ on our concept network, its first-
order nearest neighbors is denoted as $\Xi_j$. Once all its pairwise SVM classifiers are fitted on the image instances, the ensemble classifier for the given object class or image concept $C_j$ is defined as:

$$H_{C_j}(X) = \sum_{C_h \in \Xi_j} \eta_h f(C_j, C_h, X), \quad \sum_{C_h \in \Xi_j} \eta_h = 1$$

where $f(C_j, C_h, X)$ is the pairwise SVM classifier for the inter-related object classes or image concepts $C_j$ and $C_h$, and $\eta_h$ is an importance factor. By learning from all the relevant training instances (for the inter-related object classes and image concepts), our structured max-margin learning algorithm can significantly enhance the discrimination power and the generalization ability of the ensemble classifiers by combining all these pairwise SVM classifiers.

A group voting framework is developed to combine the heterogeneous pairwise SVM classifiers (that are delivered from groups of volunteers) for generating ensemble classifiers according to their reliability and confidence scores. For a given object class or image concept $C_j$, its ensemble classifier is defined as:

$$\hat{H}_{C_j}(X) = \sum_{C_h \in \Xi_j} \eta_h \sum_{l=1}^{N} \lambda_l f_l(C_j, C_h, X), \quad \sum_{C_h \in \Xi_j} \eta_h = 1, \quad \sum_{l=1}^{N} \lambda_l = 1$$

where $f_l(C_j, C_h, X)$ is the heterogeneous pairwise SVM classifier for $C_j$ and $C_h$ that is delivered from the $l$th volunteer, $\lambda_l$ is an importance factor that is strongly related with the reliability and confidence score of the $l$th volunteer’s classifier $f_l(C_j, C_h, X)$, $N$ is the total number of volunteers who deliver their heterogeneous pairwise SVM classifiers for $C_j$ and $C_h$.

Our proposed framework for collaborative classifier training can provide multiple advantages: (1) It has good scalability with the number of object classes and image concepts by leveraging both the specific abilities of groups of volunteers and the strong computational power of their unused PCs; (2) It can enhance the discrimination power of the classifiers significantly by learning from the image instances for other inter-related object classes and image concepts on the concept network, especially when the training instances available for the given object class or image concept may not be representative for large amounts of unseen test images.

4. Human-Computer Interaction and Transparency

It is true that automatic data analysis tools have freed humans from many time-consuming and labor-intensive activities [13-14]. On the other hand, human involvement (human-computer communication)
Figure 4: (a) Traditional image visualization algorithm without summarization and overlapping is visible when 2500 images are displayed; (b) Our image visualization algorithm by selecting 150 representative images to represent 28363 images without overlapping.

and human collaboration (human-human communication) still play important roles because automatic image understanding is still a difficult task for computers and human can significantly outperform computers [1-3]. Human involvement is critical for classifier training because computers must rely on human users to set goals, select alternatives if original approach fails, participate in unanticipated emergencies, and derive novel solutions [4-7]. In order to leverage human guidances and insights for collaborative classifier training, it is very important to develop new algorithms for dealing with human-computer communication more effectively.

Human-computer interaction is an important aspect of our human-centered computing system for collaborative classifier training and image understanding. Without experiencing with image instances interactively, it is very hard for volunteers to make suitable hypotheses for collaborative classifier training. Thus it is very important to develop interactive visualization frameworks to make high-dimensional image instances and their diverse visual similarity contexts to be visible, understandable and manipulable [15-18, 27-28], so that volunteers can explore the image instances interactively according to their diverse visual similarity contexts. In such interactive image exploration process, the volunteers can gain deep insights rapidly and come up new ideas to make more suitable hypotheses (i.e., select more suitable combinations of various feature subsets) for collaborative classifier training. Obviously, such an interactive image visualization and exploration tool will also support visual-based assessment of: (a) the correctness of inter-related classifiers and their decision boundaries; and (b) the effectiveness of hypotheses that are used for collaborative classifier training.
To support more effective human-computer communication, it is very important to develop new algorithms for: (1) translating human guidances (i.e., hypotheses and assessments) into computer understandable forms, so that computers can leverage such human guidances for improving classifier training; and (2) translating computer achievements (i.e., classifiers and their decision boundaries) into human understandable forms, so that humans can interactively assess both the correctness of computer achievements and the effectiveness of hypotheses for classifier training. To support more effective human-human communication, it is very important to develop new tools for knowledge and image visualization, so that humans can make their hypotheses and assessments to be visually-understandable by others and share their interesting observations and understandable assessments more intuitively.

To make such human-computer interaction and human-human communications to be transparent, it is very important to enable more effective manipulation and exploration of large-scale image instances by transforming the image instances (which are usually represented as the data points in a high-dimensional multi-modal feature space) into the forms that are more suitable to enable interactive visualization and exploration [27-28]. Obviously, such multi-modal image transformation process should faithfully retain the rich content of the image instances (i.e., their statistical properties in the high-dimensional multi-modal feature space, their diverse similarity contexts, and their correlations with the object classes and image concepts for image semantics interpretation). Some pioneering work have been done by incorporating multivariate data analysis and multi-dimensional scaling for supporting data visualization and exploration [15-18]. However, visualizing large amounts of image instances on a size-limited display screen may seriously suffer from the overlapping problem [27-28]. To reduce the visualization complexity and address the overlapping problem, it is very attractive to develop automatic summarization frameworks by selecting a small number of the most representative image instances to highlight large-scale image instances briefly. In addition, it is not a trivial task to preserve the diverse visual similarity contexts between the image instances while projecting them from a high-dimensional multi-modal feature space to a two-dimensional display space. To assist the volunteers on assessing the learning results and the hypotheses (from themselves or other volunteers on the collaboration network), it is very attractive to develop new visualization frameworks that are able to achieve automatic image summarization and similarity-preserving image projection.

When large-scale image instances come into view, most existing image visualization techniques may seriously suffer from the overlapping problem. In addition, presenting large amounts of image
instances to the volunteers may divert their attentions and cause a serious problem of huge exploration load, e.g., huge cognitive cost. In this work, an automatic image summarization framework is developed to enable interactive visualization of large amounts of image instances with less overlapping. Our interactive image visualization framework consists of multiple innovative components: (1) an image summarization algorithm to select a small set of the most representative image instances for highlighting large amounts of image instances briefly; (2) a kernel PCA (KPCA) algorithm to achieve similarity-preserving projection of the image instances from the high-dimensional multi-modal feature space to a 2D display space; (3) hyperbolic geometry to create “more spaces” for interactive image visualization and exploration by supporting change of focus.

Sparse representation has received notable attention in the multimedia and computer vision communities [24-26]. In this work, we treat image summarization as a codebook learning issue, e.g., for a given set of image instances under a given object class or image concept $C_j$, a small set of the most representative image instances are selected to briefly highlight large amounts of image instances and their diverse visual similarity contexts. In traditional sparse coding approaches [24-26], the code vectors in the codebook are usually represented as the weighted combinations of the original image instances. For image summarization application, the code vectors in our codebook (i.e., summary of a set of image instances) are the original image instances rather than their weighted combinations.

For a given image set $\Xi = \{X_1, X_2, \cdots, X_N\}$ of the given object class or image concept $C_j$, the goal of image summarization is to learn a codebook $\Omega \in \Xi$ by solving an optimization issue:

$$
\min_{\Omega, \omega} \left\{ \sum_{i \in \Xi} \|X_i - \Omega \omega_i\|^2_2 + \lambda \|\omega_i\|_1 \right\}
$$
where $\ell_1$-norm is used for enforcing sparsity, $\lambda$ is the parameter to balance the representation fidelity and sparsity of the solution, $\omega = \{\omega_1, \cdots, \omega_N\}$ is the set of weights.

An iterative algorithm is developed for learning the codebook $\Omega$ (summary of the given image set $\Xi$): (a) all the image instances in $\Xi$ are first treated as the initial code vectors in the codebook $\Omega$; (b) the sub-optimal set of the weights $\omega$ are learned via linear programming; (c) when the initial set of the weights $\omega$ is obtained, the sub-optimal codebook $\Omega$ is obtained via quadratically constrained quadratic programming; (d) when the sub-optimal codebook is determined, go back step (a) for searching a better set of the weights $\omega$; (e) the above processes are performed repeatedly until the number of non-zero weights $\omega$ reaches a pre-defined threshold and the image instances with non-zero weights $\omega$ are selected as the most representative instances. As shown in Fig. 4, presenting a small number of the most representative instances to the volunteers can significantly reduce their exploration load (i.e., cognitive cost) while allowing them to gain deep insights rapidly.

The most representative image instances are projected onto a hyperbolic plane by using kernel PCA to preserve their diverse visual similarity contexts precisely, and the kernel PCA (KPCA) is achieved by solving the eigenvalue equation [20]. There are many potential projections from a high-dimensional multi-modal feature space to a 2D display space, the optimal KPCA-based projection of the image instances is selected by minimizing the cumulative differences between their original similarity distances in the high-dimensional multi-modal feature space and their Euclidean distances on a 2D display space. Such KPCA-based projection can faithfully preserve the original similarity contexts between the most representative image instances.

Poincaré disk model is used to map the most representative image instances on the hyperbolic plane onto a 2D display coordinate [18, 20]. Poincaré disk model can map the entire hyperbolic space onto an open unit circle, and produce a non-uniform mapping with “more display space”. Such Poincaré mapping can easily support change of focus to allow volunteers to explore the image instances interactively according to their diverse visual similarity contexts, so that they can change the presentation and visualization of the image instances interactively for evaluating the correctness of the decision functions and assessing the effectiveness of the hypotheses that are used for collaborative classifier training. By making the hypotheses from different volunteers to be visible and understandable (some examples are given in Fig. 5), our interactive image visualization framework can provide a good communication environment to interactively handle the huge uncertainty of human guidances (e.g.,
different volunteers may make significantly different hypotheses for the same classifier training task and some of these hypotheses may not be correct or effective at all). Through change of focus, the volunteers can quickly change their views of the image instances and directly see what is missing, what is expected, what is unexpected, and what is conjectured, so that they can infer the alternative solutions for collaborative classifier training.

Our interactive visualization framework can allow the volunteers to explore the image instances under all possible combinations of the feature subsets, so that they can assess the effectiveness of different combinations of the feature subsets interactively. In such interactive image exploration process, the volunteers can rapidly gain deep insights from the image instances, and they can also change the hypotheses by selecting different combinations of the feature subsets according to their personal observations. As shown in Fig. 5, the volunteers can also consider multiple competing hypotheses simultaneously, so that they can carry out multiple alternative solutions (i.e., train multiple classifiers alternatively) for the same task. Thus some hypotheses (i.e., various combinations of the feature subsets), which may not be expected when the process for classifier training was commenced, can be suggested by our structural max-margin learning algorithm and be evaluated interactively by the volunteers. Given the new hypotheses (i.e., human guidances), our structural max-margin learning algorithm can train new classifiers for automatic object and concept detection. By involving the volunteers in the processes for interactive hypotheses making and collaborative classifier training, our proposed research can significantly boost the capabilities for both human beings and machine learning techniques for collaborative image understanding. By making the hypotheses to be visible and understandable, the volunteers can interactively assess the effectiveness and correctness of the hypotheses from themselves and others.

5. Conclusion

In this paper, we have proposed a human-centered computing framework for supporting collaborative image understanding by leveraging the essential characteristics of both humans and computers. By involving massive numbers of volunteers and millions of unused computers worldwide for collaborative classifier training, our proposed research may provide an unique and innovative way for tackling the challenging issue of automatic image understanding collaboratively. Our future research will focus on developing an online collaboration environment and inviting multimedia researchers to join our project.
References


