Learning Feature Hierarchies: A Layer-Wise Tag-Embedded Approach

Zhaoquan Yuan, Changsheng Xu, *Fellow, IEEE*, Jitao Sang, Shuicheng Yan, *Senior Member, IEEE*, and M. Shamim Hossain, *Senior Member, IEEE*

***Abstract*—Feature representation learning is an important and fundamental task in multimedia and pattern recognition research. In this paper, we propose a novel framework to explore the hierarchical structure inside the images from the perspective of feature representation learning, which is applied to hierarchical image annotation. Different from the current trend in multimedia analysis of using pre-defined features or focusing on the end-task "flat" representation, we propose a novel layer-wise tag-embedded deep learning (LTDL) model to learn hierarchical features which correspond to hierarchical semantic structures in the tag hierarchy. Unlike most existing deep learning models, LTDL utilizes both the visual content of the image and the hierarchical information of associated social tags. In the training stage, the two kinds of information are fused in a bottom-up way. Supervised training and multi-modal fusion alternate in a layer-wise way to learn feature hierarchies. To validate the effectiveness of LTDL, we conduct extensive experiments for hierarchical image annotation on a large-scale public dataset. Experimental results show that the proposed LTDL can learn representative features with improved performances.**

***Index Terms*—Auto-encoder, deep learning, hierarchical feature learning, social tags.**

I. Introduction

THE performance of an artificial intelligence system is highly dependent on the choice of data representation (or features) [1], for different representations may entangle and hide more or less the different explanatory factors of variations behind the data. Recently, feature representation has become the focus of many application-related work [2], [3].

In the recent few years, many methods which aim at obtaining representation features have been developed, ranging from hand-crafted features such as SIFT [4] and HOG [5], to auto-learned features which are based on machine learning methods such as PCA [6], ICA [7], sparse coding [8], etc. These representation methods have achieved great success, and are widely used in many research areas including computer vision and multimedia content analysis. However, these existing methods focus on the end tasks and the feature representation is "flat", and the structured representation is largely ignored, which limits the ability to analyze and handle more sophisticated tasks, for the hierarchical structure of representation is not explored enough.

Different from existing methods, in this paper, we aim to learn the hierarchical features which represent an image with multi-level structures. Different levels of the hierarchical features convey different semantics. Our motivations are threefold. First, the hierarchical feature is more consistent with the inherent characteristic of human cognition. Research findings in cognitive science show that the human visual system follows a similar hierarchical structure, with higher levels representing more complex features [9], [10]. Developing a similar human cognitive system has always been the pursuit of the artificial intelligence field. Second, the hierarchical feature represents objects in a more exquisite way. The hierarchical feature organizes the representation into multiple levels according to the corresponding semantic levels, which is more sophisticated and refined for representation. Third, for the images with contextual information, the semantics are essentially hierarchical. For example, Fig. 1(a) shows an image with semantic tags: *animal, plant, cat, dog*, and *flower*. Obviously, these semantic tags are not in the same level, but compose a hierarchy. From above observation, we see that learning feature hierarchies is a promising and important research topic for both representation learning and image-related multimedia analysis.

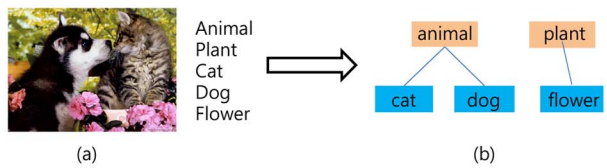


Fig. 1. Image and associated social tags (a), and the corresponding tag hierarchy (b).

Manuscript received December 22, 2014; revised March 11, 2015; accepted March 11, 2015. Date of publication March 27, 2015; date of current version May 13, 2015. This work was supported in part by the National Basic Research Program of China under Grant 2012CB316304, by the National Natural Science Foundation of China under Grant 61225009, Grant 61373122, Grant 61303176, Grant 61402479, Grant 61328205, and Grant 61432019, by the Beijing Natural Science Foundation under Grant 4131004, by the Singapore National Research Foundation under its International Research Centre@Singapore Funding Initiative and administered by the IDM Programme Office, and by the Deanship of Scientific Research at King Saud University through the international research group Program IRG 14-18. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Cees Snoek.

Z. Yuan, C. Xu, and J. Sang are with the National Lab of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China (e-mail: zqyuan@nlpr.ia.ac.cn; csxu@nlpr.ia.ac.cn; jtsang@nlpr.ia.ac.cn).

S. Yan is with the Department of Electrical and Computer Engineering, National University of Singapore, Singapore 119077 (e-mail: eleyans@nus.edu.sg).

M. S. Hossain is with the SWE Department, College of Computer and Information Sciences, King Saud University, Riyadh 11543, Saudi Arabia (e-mail: mshossain@ksu.edu.sa).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TMM.2015.2417777

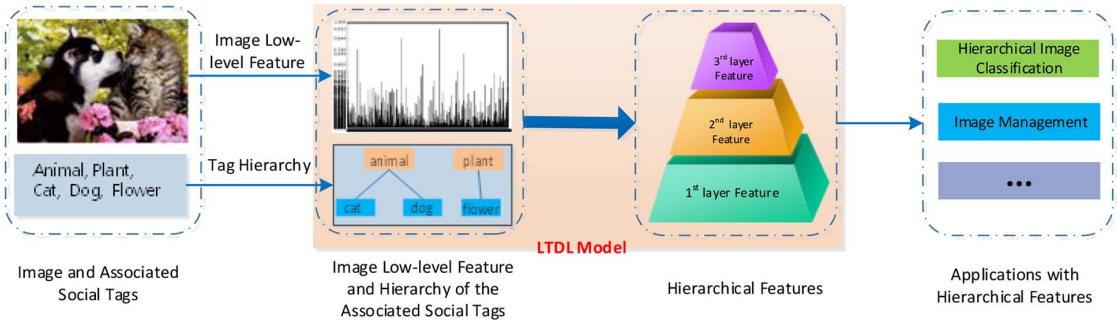


Fig. 2. Illustration of the feature hierarchy learning process.

However, it is a challenging task to gain hierarchical features according to their corresponding semantics. In the feature hierarchy, the relationship between the feature layers needs to be explored. Usually, the features in the lower layer compose, or are integrated into the features in the higher layer [11], [12], which results in the combination of lower-level features into more abstract features and the formation of the feature hierarchies. It is difficult to achieve this goal for regular shallow models due to the limitation of local learning and curse of dimensionality [13]. Fortunately, the recent research of deep learning [14], [15], which attempts to model high-level abstraction in data by using architectures composed of multiple non-linear transformations, provides the potential for achieving our goal. However, it is still a challenging problem to use the current deep learning based methods for feature hierarchy learning. Most of the existing deep learning methods only utilize the visual content information of the images, and the learned feature hierarchies are implicit (i.e., the semantics for which the feature is responsible are unclear). Besides, although the layer-wise learning mechanism of deep learning fits well to the hierarchy construction, the existing deep methods train the parameters using layer-wise unsupervised pre-training and global fine-tuning, and there is no supervised information between the layers. This leads to the fact that the relationship between feature layers is implicit rather than explicit and the property of the hierarchy is weak.

Information in the real world usually comes from multiple channels which are usually complementary to each other. Some related research work with multiple modality information of the images to improve the performance [16], [17] have proven this point. In the social media platform, the images are usually associated with social tags as shown in Fig. 1(a) and these tags indicate the semantics of the image to some degree. We believe that these tags can provide complementary information for the images and are helpful for our feature learning task. Furthermore, the tags usually constitute a tree-like hierarchical semantic to a certain extent and can be converted into a hierarchy by existing tools, such as WordNet [18] which groups words into sets of synonyms and records different semantic relations between them. Fig. 1 shows an image and the associated tags from Flickr in (a), and the converted tag hierarchies in (b). We argue that the hierarchical structure of the tags, which can be considered as a special kind of modality of an image, can be utilized to improve the feature hierarchy learning in our task. On one hand, the semantic tags in each level can provide the supervised information for each layer of features in the feature hierarchy, which makes the meaning of the feature explicit. On the other hand, it can provide a new fusion mechanism for multiple modalities in the different semantic levels.

In this paper, we propose to learn feature hierarchies by utilizing both the visual content information and the tag hierarchy information, as illustrated in Fig. 2. Given the images and the associated social tags, our system learns the feature hierarchies which consist of multiple levels of feature representations. Each layer of the feature corresponds to a different semantic level. The higher the layers are, the more abstract and general the features will be. We propose a novel Layer-wise Tag-embedded Deep Learning (LTDL) model to integrate the social tag information into the visual content in a layer-wise way. In our LTDL model, the concept hierarchy is not just used to organize training images. It is integrated into feature learning from two aspects. Firstly, to handle the limitation of no inter-layer supervised information in the conventional deep methods, we utilize the hierarchical social tags to restrict the inter-layer relationship by learning the parameters of the deep network with layered supervised methods. Each layer of the tag hierarchy is used for supervised learning the corresponding feature layer. Secondly, each tag layer is fused with current feature layer for learning feature in the higher layer, which is a multimodal fusion step. During the whole learning process, the supervised training and the multi-modal fusion alternate in a layer-wise way.

Based on the LTDL model, we can complement the missing tags for the images, which can be considered as a kind of hierarchical annotation system to annotate the images in different semantic levels. We evaluate the proposed LTDL on a large-scale corpus of over one million images. The experimental results have demonstrated the advantage of the proposed method.

The contributions of our work can be summarized as follows.

- We are the first to propose learning features of social images in a hierarchical way, which is more refined representation than the conventional "flat" features.
- We propose a novel LTDL model that fuses the social tag hierarchy and the visual content information seamlessly to learn hierarchical feature representation.
- Based on the feature hierarchy of the images, we develop a hierarchical annotation system and conduct a series of experiments on a large-scale image dataset to validate the effectiveness of our work.

The rest of the paper is organized as follows. In Section II, we review the state-of-the-arts most related to our research. In Section III, we present our proposed LTDL model. Experi-

mental results and analysis are reported in Section IV. We conclude the paper with future work discussions in Section V.

II. Related Work

In this section, we review the work most related to our research, including feature learning, concept hierarchy-based multimedia analysis, and hierarchical models.

A. Feature Learning

Feature learning is a classical problem in machine learning, and extensive efforts have been devoted to this research area. In this section, we briefly review three kinds of feature learning methods: conventional feature learning, hierarchical feature learning, and cross-modality feature learning.

Conventional feature learning is based on conventional machine learning methods, and to learn the latent representation. One kind of methods in this category aims to find a low-rank approximation for the raw features, including principal component analysis (PCA) [6], independent component analysis (ICA) [7], etc. They map the raw data to a lower-dimensional representation based on the assumption that the data lie (approximately) in an underlying low-dimensional linear subspace. Another kind of methods is based on the single-layer network, e.g., RBM, auto-encoder, and K-means clustering. Reference [19] demonstrated that large numbers of hidden nodes and dense feature extraction are critical for achieving high performances of these methods. In addition, the topic model [20] can also be regarded as a solution of feature learning, which aims to discover the abstract "topics" that occur in a collection of raw data and take the "topics" distributions as the semantic representation of the raw data.

Some pioneering work for hierarchical feature has been done, especially in dictionary learning and sparse coding [8]. References [21] and [22] proposed a joint dictionary learning algorithm to exploit the visual correlation within a group of visually similar object categories for dictionary learning where a commonly shared dictionary and multiple category-specific dictionaries are accordingly modeled. Reference [23] proposed a novel dictionary learning method by taking advantage of hierarchical category correlation. For each internode of the hierarchical category structure, a discriminative dictionary and a set of classification models are learnt for visual categorization, and the dictionaries in different layers are learnt to exploit the discriminative visual properties of different granularity.

Cross-modality feature learning focuses on combination of data in different modality data space. [24] learned cross-modal correlation between visual and auditory feature spaces, and treated such correlation as complementary information for clustering on image-audio dataset. Reference [25] proposed a coupled linear regression framework to deal with the problem of cross-modal matching. Reference [26] presented a cross-modal approach for extracting semantic relationships between concepts using tagged images based on canonical correlation analysis (CCA).

From the perspective of the depth of the model architecture, the above methods are shallow models. Shallow models suffer from the curse of dimensionality, and have limited capability in learning the distributed representation in complex situations [27].

Recently, a novel kind of deep learning framework [28], which models the learning task by using deep architectures composed of multi-layer nonlinear modules, rises up to the challenges and provides a proper tool for automatic feature learning. Typical deep models include Deep Belief Nets [29], Deep Auto-encoder [30], Conventional Neural Network [31], etc. Considering the strong power of feature learning from the visual content, we integrate the concept hierarchy into the deep model [30] to learn the feature hierarchy.

B. Concept Hierarchy

Concept hierarchy organizes the concepts into a tree structure according to their semantics, where each node represents a semantic concept, such as WordNet [18] and LSCOM [32]. Recent work on exploiting the concept hierarchy for multimedia analysis involves a wide kinds of applications [33]–[36]. For example, [33] proposed a semantic hierarchical classifier that uses the semantics of image labels to extract knowledge about the inter-class relationships and integrates it into the visual appearance learning procedure. Reference [37] processed the input data by applying a hierarchy-based heuristics for feature selection and feature aggregation. Most of the existing work on using the concept hierarchy for image/video retrieval focused on semantic similarity computation between query and images/videos by integrating the concept hierarchical information. In the method proposed by [35], given two images, their visual nearest neighbor images are first found, and then their semantic distance is computed as a distance between the concepts of their neighbors. Reference [36] developed a hierarchical bilinear similarity function for image retrieval.

Different from the existing methods which exploit the concept hierarchy for end-task multimedia analysis, we embed the concept hierarchy information into the visual content to learn the hierarchical feature representation. To the best of our knowledge, it is the first time that concept hierarchy is used for learning feature hierarchies.

C. Hierarchical Models

There are a number of hierarchical models for visual recognition and multimedia analysis. Compared with the conventional "flat" structure model, the hierarchical models handle data analysis with more than one level of parameters [38]. According to the learning ways of the parameters, hierarchical structured models can be divided into two categories: hierarchical Bayesian models and discriminative hierarchical models, both of which are widely used for multimedia analysis.

Hierarchical Bayesian models represent the data within the Bayesian framework by defining the variables in the higher layer as prior knowledge of the variables in the lower layer. Typical hierarchical Bayesian models include various hierarchical topic models [39] which address the problem of learning topic hierarchies from data. Reference [40] utilized the Chinese restaurant process (CRP) for multi-modal visual dictionary learning. Similarly, [41] proposed a hierarchical topic model based multi-modal framework for web video faceted subtopic

retrieval. Reference [42] explored the hierarchical topic structure in the retrieved video collection and presented users with videos organized into semantic clusters.

Different from the hierarchical Bayesian models, discriminative hierarchical models represent the data by directly transforming the variables in the lower layer to those in the higher layer rather than defining the priors. Various transforming methods such as CRF and MRF can be used. Reference [43] proposed a hierarchical CRF model to deal with the problem of labeling images of street scenes with several distinctive object classes. Analogously, [44] and [45] utilized hierarchical CRF for non-rigid object detection and object class image segmentation respectively. Reference [46] and [47] utilized the hierarchical MRF for sonar picture segmentation and multi-object tracking respectively.

Note that the deep learning model can also be considered as a hierarchical model, since the deep architecture includes multiple parameters and models the data into a hierarchy. They are widely used for visual recognition [48], image retrieval [49], etc.

The existing hierarchical models focus on modeling the visual content information while the corresponding hierarchical information of tags/concepts is ignored. Therefore, these methods cannot be directly used in our task.

III. Learning Feature Hierarchies

In this section, we propose a novel Layer-wise Tag-embedded Deep Learning (LTDL) model to seamlessly combine the visual content and the associated tag hierarchy information to gain more representative hierarchical features. Meanwhile, we utilize the hierarchical features learned by the LTDL model for the hierarchical image annotation task.

A. Layer-Wise Tag-Embedded Deep Learning Model

A tag hierarchy (taxonomy) G is a forest whose trees are defined over a set of tags, and a multi-label $\mathbf{y} \in \{0,1\}^M$ is said to respect a tag hierarchy if and only if $\mathbf{y}$ is the union of one or more paths in G, where each path starts from a root but need not terminate on a leaf [50]. For an image instance $\mathbf{x}$, its multi-label is $\mathbf{y} = (y_1, \ldots, y_i, \ldots, y_M) \in \{0,1\}^M$ which is any subset of the whole tag set $\{1, 2, \ldots, M\}$, where $i \in \{1, \ldots, M\}$ belongs to the multi-label of $\mathbf{x}$ if and only if $y_i = 1$.

Given the image visual content feature $\mathbf{x}$, we divide the associated multi-label $\mathbf{y}$ into K levels according to the tag hierarchy G to make $\mathbf{y} = \{\mathbf{y}^{(L-K+1)}, \ldots, \mathbf{y}^{(L)}\}$ with K $(K < L)$ levels from bottom to up. Our goal is to learn a hierarchical feature representation $\mathbf{h} = \{\mathbf{h}^{(L-K+1)}, \ldots, \mathbf{h}^{(L)}\}$ with the feature in $\mathbf{h}^{(l)}$ corresponding to the tags in level $\mathbf{y}^{(l)}$, where $L - K + 1 \leq l \leq L$. Since in the tag hierarchy, the tags in the higher levels are more abstract than those in the lower levels, the features in the higher layers are also more abstract than those in the lower layers in the final learned feature hierarchy.

For convenience, the symbols used in this model are summarized in Table I. Since the visual content and the tag hierarchy are with different modalities and different intrinsic data structures in spaces, it is challenging to fuse them into a latent hierarchical feature. Firstly, the low-level visual feature is a "flat" vector for the image and there is no multiple representation while the tag hierarchy represents the image with multiple levels of semantics. Secondly, it is difficult to make the features in each layer in the learned feature hierarchy correspond to the semantic tags in the same level, which involves the problem of feature interpretation for each layer. Thirdly, it is challenging to design a reasonable fusion mechanism of two kinds of structured information even if both of them are hierarchical. To the best of our knowledge, there is no existing related work.

TABLE I
Summary of the Symbols

Symbol	Description
L	number of layers of deep architecture (except the lowest input layer)
K	number of layers of the tag hierarchy
$\mathbf{x}(\mathbf{h}^{(0)})$	low-level raw feature vector of an image
$\mathbf{h}$	feature hierarchy with K levels
$\mathbf{y}$	tag hierarchy with K levels
$\widetilde{\mathbf{h}}$	corrupted vector from $\mathbf{h}$
ν	corruption rate
$\mathbf{z}^{(l)}$	reconstruction vector for layer l
$Loss(\,)$	loss function
$\mathbb{H}(\mathcal{B}_{\mathbf{x}}\|\|\mathcal{B}_{\mathbf{z}})$	cross-entropy between $\mathcal{B}_{\mathbf{x}}$ and $\mathcal{B}_{\mathbf{z}}$
$\mathbf{h}^{(l)}$	feature vector of l-th layer in the deep architecture
$\mathbf{h}^{(l)}_{mer}$	merging of $\mathbf{h}^{(l)}$ and $\mathbf{y}^{(l)}$
$\mathbf{W}^{(l)}_h$	weight parameter between layer l and layer $l+1$ in the feature network
$\mathbf{W}^{(l)}_y$	weight parameter between layer l in the tag hierarchy and layer $l+1$ in the feature network
$\mathbf{b}^{(l)}$	bias parameter of layer l
$s(\mathbf{x})$	sigmoid function: $s(\mathbf{x}) = \frac{1}{1+e^{-\mathbf{x}}}$
$D^{(l)}$	dimension of the vector $\mathbf{h}^{(l)}$
$M^{(l)}$	number of the tags in layer l in the tag hierarchy
$\mathcal{B}_{\mathbf{z}}$	Bernoulli distribution with mean $\mathbf{z}$
N	number of the training samples
M	number of the whole tags

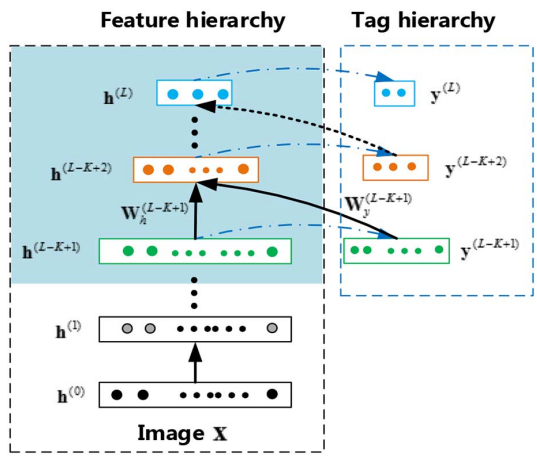


Fig. 3. Layer-wise tag-embedded deep learning model.

Towards the challenges discussed above, we design a novel LTDL model which is shown in Fig. 3. In our model, we construct a hierarchical structured model with L layers, where the variables in the top K layers correspond to the feature hierarchies in the K layers. In our model, the visual content information and the corresponding tags are fused in a layer-wise way.

Deep learning models, which learn the feature hierarchy from visual content by composition and decomposition of the lower-level features, fit well to our task. On one hand, from the view of the structure, they learn the visual features in the same layer-

wise way, and obtain the hierarchical structure; on the other hand, deep learning models have strong power of feature extraction [1], [27], [51]. Among these deep models, deep auto-encoder is an effective and easy-training model [52], based on which we construct the visual content feature hierarchy. Different from the mainstream training mechanism with two separate stages of pre-training and fine-tuning in the current deep auto-encoder model, our method trains the architecture by alternating the two stages.

Note that the deep architecture has L layers, which is larger than the number of levels in the tag hierarchy. If we set L equal to K, the features in the lowest layer correspond to the tags $\mathbf{y}^{(L-K+1)}$. However, we argue that it is difficult to learn semantic representation from the low-level raw features directly. Hence, in our model, from the low-level raw feature space to the feature space corresponding to the semantic tags $\mathbf{y}^{(L-K+1)}$, there are some layers of non-linear mappings, and only this can result in the consistence of the feature $\mathbf{h}^{(L-K+1)}$ with the semantic tags $\mathbf{y}^{(L-K+1)}$.

Algorithm 1: Layer-Wise Tag-Embedded Deep Learning Model

Input:
Low level features of images $\mathbf{x}_1, \ldots, \mathbf{x}_N$;
Number of deep network layers L;
Number of tag hierarchy layers K;
Tag hierarchy $\mathbf{y}^{(L-K+1)}, \ldots, \mathbf{y}^{(L)}$;
Initial weights and bias parameters $\{\mathbf{W}\}$, $\{\mathbf{b}\}$;
Output:
The missing hierarchical tags $\mathbf{y}^{(L-K+1)}, \ldots, \mathbf{y}^{(L)}$

for *each training sample* $\mathbf{x}_n$ **do**
  for *each layer* $\mathbf{h}^{(l)}$ *from* $\mathbf{h}^{(1)}$ *to* $\mathbf{h}^{(L-K+1)}$ **do**
    Greedy layer-wise pre-training by learning a denoising auto-encoder at a time;
  for *each layer* $\mathbf{h}^{(l)}$ *from* $\mathbf{h}^{(L-K+1)}$ *to* $\mathbf{h}^{(L)}$ **do**
    Supervised fine-tuning the parameters from the layer $\mathbf{h}^{(0)}$ to layer $\mathbf{h}^{(l)}$;
    Learning the multimodal auto-encoder between the $\mathbf{h}^{(l)}$, $\mathbf{y}^{(l)}$, and $\mathbf{h}^{(l+1)}$;

From the low-level feature $\mathbf{x}(\mathbf{h}^{(0)})$ in the lowest layer to the layer $\mathbf{h}^{(L-K+1)}$, each two adjacent layers constitute an auto-encoder. In order to make the model more robust, we take the denoising version of the auto-encoder [30]. For the adjacent layers $\mathbf{h}^{(l-1)}$ and $\mathbf{h}^{(l)}(1 \le l \le L-K+1)$, a fixed number νd of components are chosen randomly, and their values are forced to 0, while the others are left untouched and we get the corrupted vector $\widetilde{\mathbf{h}}^{(l-1)}$. Then we transform it to get the $\mathbf{h}^{(1)}$ representation

$$\mathbf{h}^{(1)} = s\left(\mathbf{W}_h^{(l-1)}\,\widetilde{\mathbf{h}}^{(l-1)} + \mathbf{b}^{(l)}\right). \tag{1}$$

Then the latent representation $\mathbf{h}^{(l)}$ is mapped back to a "reconstructed" vector $\mathbf{z}$

$$\mathbf{z}^{(l-1)} = s\left(\mathbf{W}_h'^{(l-1)}\mathbf{h}^{(l)} + \mathbf{b}'\right) \tag{2}$$

where $\mathbf{W}_h'^{(l-1)}$ is the transposition of $\mathbf{W}_h^{(l-1)}$. The parameters $\theta = (\mathbf{W}_h^{(l-1)}, \mathbf{b}^{(l)}, \mathbf{b}')$ are optimized to minimize the reconstruction cross-entropy

$$\begin{aligned}&\mathbb{H}(\mathcal{B}_{\mathbf{h}^{(l-1)}} \| \mathcal{B}_{\mathbf{z}}) \\ &= -\sum_{k=1}^{D^{(l-1)}} \left[h_k^{(l-1)} \log z_k^{(l-1)} + (1 - h_k^{(l-1)}) \log(1 - z_k^{(l-1)})\right].\end{aligned} \tag{3}$$

After performing this process in a layer-wised unsupervised way from a low-level feature $\mathbf{x}(\mathbf{h}^{(0)})$ in the lowest layer to the layer $\mathbf{h}^{(L-K+1)}$, we finally gain the feature vector $\mathbf{h}^{(L-K+1)}$, which can be considered as the preliminary latent representation of the visual content. It then can be used for later fusion with the tag hierarchy.

Towards the challenges of feature interpretation for each layer and the difficulty of fusion of visual content and tag hierarchy information, we take two steps in each layer alternately: *supervised training* and *multimodal fusion*. Next, we introduce the two steps for hierarchical feature learning and explain the ideas behind the operations.

The latent presentation from the unsupervised pre-training is completely based on the mapping and reconstruction operations, which is more abstract than the low-level visual feature. However, on one hand, this is not enough for explicit feature hierarchy in our task. It is unclear to which semantic level in the tag hierarchy the current feature layer corresponds if only the low-level visual information is used, and the problem of feature interpretation cannot be addressed. On the other hand, the contextual tag information can provide a way to refine the parameters of the nets below the current layer which results in more representative features.

Based on the ideas discussed above, in our model, we firstly use tags $\mathbf{y}^{(l)}$ $(L-K+1 \le l < L)$ to fine-tune the parameters from layer 0 to layer l, and the cross-entropy loss is used

$$Loss\left(\{\mathbf{W}, \mathbf{b}\}\right) = -\sum_{n=1}^{N}\sum_{k=1}^{M^{(l)}} y_{nk}^{(l)} \ln p_{nk}^{(l)} \tag{4}$$

where $p_{nk}^{(l)}$ is the predicted tag value of the k-th dimension of the n-th sample from the feature $\mathbf{h}^{(l)}$. Based on the fine-tuned parameters, the refined latent feature representation $\mathbf{h}^{(l)}$ can be gained by multiple non-linear mapping from bottom to up, and we can consider that $\mathbf{h}^{(l)}$ corresponds to the semantic representation of $\mathbf{y}^{(l)}$.

The semantic information is hierarchical and we fuse the visual content and the tag information in a layer-wise way. Hence after gaining the current feature representation $\mathbf{h}^{(l)}$, we next try to gain the higher representation $\mathbf{h}^{(l+1)}$. Different from the conventional deep methods where only $\mathbf{h}^{(l)}$ is utilized, we combine the feature $\mathbf{h}^{(l)}$ and the tag $\mathbf{y}^{(l)}$ to learn a multi-modal deniosing auto-encoder, which learns a probabilistic density of $\mathbf{h}^{(l+1)}$ over

the space of multi-modal inputs. Given $\mathbf{h}^{(l)}$ and $\mathbf{y}^{(l)}$, the values of $\mathbf{h}^{(l+1)}$ are computed by

$$\mathbf{h}^{(l+1)} = s\left(\mathbf{W}_h^{(l)}\,\widetilde{\mathbf{h}}^{(l)} + \mathbf{W}_y^{(l)}\,\mathbf{y}^{(l)} + \mathbf{b}^{(l+1)}\right). \tag{5}$$

The parameters are optimized to minimize a modified reconstruction cross-entropy

$$\begin{aligned}&\mathbb{H}(\mathcal{B}_{\mathbf{h}^{(l)}}, \mathcal{B}_{\mathbf{y}^{(l)}} \| \mathcal{B}_{\mathbf{h}^{(l+1)}}) \\ &= -\sum_{k=1}^{D^{(l)}}\left[h_k^{(l)} \log z_{\mathbf{h}k}^{(l)} + (1-h_k^{(l)})\log(1-z_{\mathbf{h}k}^{(l)})\right] \\ &\quad -\sum_{k=1}^{M^{(l)}}\left[y_k^{(l)} \log z_{\mathbf{y}k}^{(l)} + (1-y_k^{(l)})\log(1-z_{\mathbf{y}k}^{(l)})\right]\end{aligned} \tag{6}$$

where $\mathbf{z}_h^{(l)}$ is the "reconstructed" vector for the $\mathbf{h}^{(l)}$, and $\mathbf{z}_y^{(l)}$ is the "reconstructed" vector for the $\mathbf{y}^{(l)}$. They are parts of the vector $\mathbf{z}^{(l)}$. We believe that this multi-modal fusion is valuable. Firstly, the two modalities typically carry different kinds of information, and they are usually with complementary correlational structure [16]. The multi-modal fusion can help the higher layer feature $\mathbf{h}^{(l+1)}$ to carry clearer semantic information. Secondly, and most importantly, when we optimize the parameters in (6), $\mathbf{W}_\mathbf{h}^{(l)}$ and $\mathbf{W}_\mathbf{y}^{(l)}$ are optimized simultaneously, and the information in the tags will be transmitted to $\mathbf{W}_\mathbf{h}^{(l)}$, which is the target parameter to be learned in our task. In addition, as we consider $\mathbf{h}^{(l)}$ and $\mathbf{y}^{(l)}$ in the same semantic layer, the merger is reasonable.

After the fusion of feature and tags, we perform the step of supervised training, and use the tags $\mathbf{y}^{(l+1)}$ to fine-tune in a supervised way the parameters which can gain the refined latent feature $\mathbf{h}^{(l+1)}$.

The two steps alternate until the top layer of the network is reached. In this way, we achieve three goals: 1) The visual content information and the tag information are fused in a layer-wise way, and this process is natural and in accordance with the current feature learning mechanism; 2) The final features in multiple layers constitute a feature hierarchy. This hierarchical feature is more representative than regular "flat" features; 3) The feature in each layer is responsible for the corresponding tags in the current layer, which makes the hierarchy property clearer.

B. Modeling Task

In many cases, there is only visual information of the images and the associated tags are absent. For these images, our LTDL model combining the associated tag hierarchy can generate missing tags in a hierarchical way, which can be viewed as hierarchical image annotation.

The modeling task of hierarchical image annotation based on the tag hierarchy is defined as follows: for an image instance $\mathbf{x}$, our goal is to predict its multi-label $\widehat{\mathbf{y}}$, and meanwhile to meet conditions that every data instance is labeled with a (possibly empty) set of tag nodes, and whenever an instance is labeled with a certain node i, it is also labeled with all the nodes on the path from the root of the tree down to the node i.

Based on the trained deep architecture, for an image with the low-level feature $\mathbf{x}$, we firstly map it into the feature space $\mathbf{h}^{(L-K+1)}$ through multiple non-linear transformation from the layer $\mathbf{h}^{(0)}$ to $\mathbf{h}^{(L-K+1)}$ in an unsupervised layer-wise way. Then for the learned features $\mathbf{h}^{(l)}(L-K+1 \le l \le L)$ in each layer, we can generate the corresponding tags $\widehat{\mathbf{y}}^{(l)} = (\widehat{y}_1^{(l)}, \ldots, \widehat{y}_i^{(l)}, \ldots, \widehat{y}_{M^{(l)}}^{(l)})$. The probability of the i-th component $y_i^{(l)}$ is predicted by

$$P\left(\widehat{y}_i^{(l)} = 1\right) = \frac{\exp\left(\mathbf{W}_i^T \mathbf{h}_i^{(l)}\right)}{\sum_j \exp\left(\mathbf{W}_j^T \mathbf{h}_j^{(l)}\right)} \tag{7}$$

where $\mathbf{W}_i$ denotes the weights between the tag $\widehat{y}_i^{(l)}$ and the feature $\mathbf{h}^{(l)}$, and it is trained in the supervised fine-tuning stage by (4). Based on the prediction probabilities, we can get the annotation tags $\widehat{\mathbf{y}}$ for each layer by selecting the tags with maximum probability for each layer. Note that, in the prediction process, since the tags are missing, the multi-modal fusion step in the training set is not conducted in the test set.

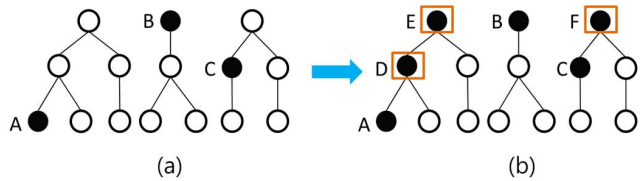


Fig. 4. Annotation results from the prediction probability (a), multi-path labelling, and partial-path labelling (b).

In order to make the annotation results meet the condition that if an instance is labeled with the node i, it should also be labeled with the parent of node i $PAR(i)$, we generate the hierarchical tags by labeling the parents of labeled nodes recursively until reaching the top layer. Fig. 4 shows an example for our method. Fig. 4(a) shows the annotation results from the prediction probability, and the nodes A, B, C are labeled for the instance. According to the condition, we also label the node D (parent of node A) and node E (parent of node D), and it is in the same way for node C, which is shown in Fig. 4(b). This method is named multi-path labeling and partial-labeling in some related literature [53], for the instance can be labeled with nodes belonging to more than one path or a path that does not end on a leaf in the forest. Note that other than multi-path labeling and partial-labeling, it seems that the method of finding the path in the hierarchy with the highest score is alternative choice. However, this method will transfer the hierarchical classification into regular classification, where each class is a path, and the classic H-loss evaluation is not applicable for it.

Different from the conventional "flat" image annotation methods, our model can generate the tags in different levels according to different latent features and the tag hierarchy, which makes the annotation more refined.

Note that the hierarchical feature learning is the core focus in this paper, and the hierarchical feature learned in the model can be applied to a range of multimedia tasks. Hierarchical tag generation is one of the most common and important tasks. As the features are utilized in different ways in different application circumstances, the modeling task is application-dependence or data-dependence, and the configuration of the prediction needs to be designed carefully.

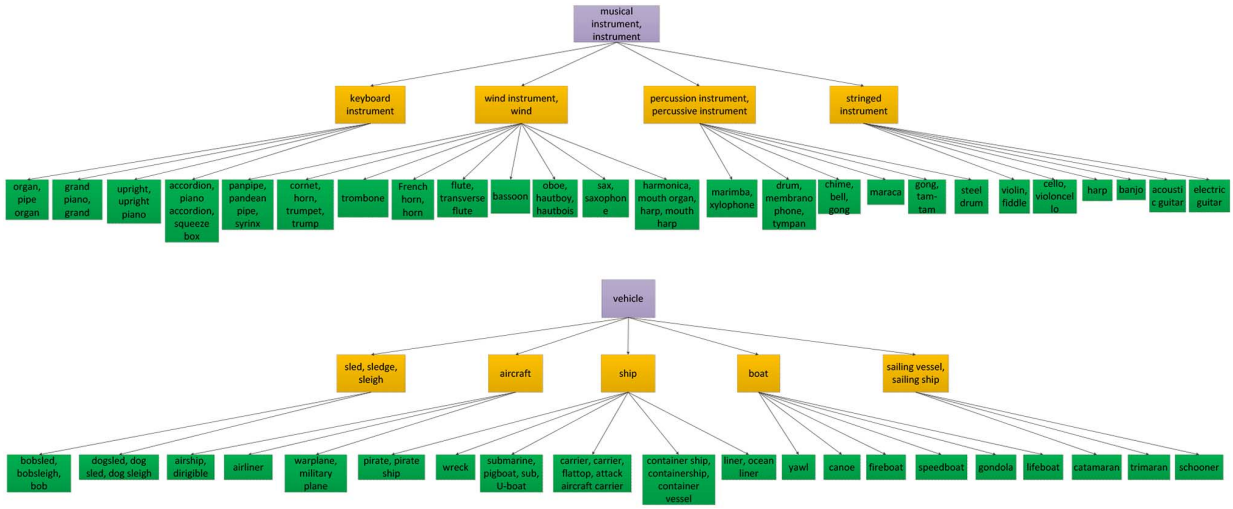


Fig. 5. Two examples of tag hierarchy of our dataset.

The proposed Layer-wise Tag-embedded Deep Learning (LTDL) model is summarized in Algorithm 1.

IV. Experiments

In this section, we evaluate the proposed LTDL model on a large-scale image dataset for hierarchical image annotation. In the experiments, we use two kinds of metrics for the evaluation: multi-class AUC (Area Under an ROC Curve) [54] which is directly based on the prediction probabilities, and the hierarchical loss which is based on the explicit annotation results.

A. Dataset

We conduct a series of experiments on ImageNet [55], which is a large-scale ontology of images organized according to the WordNet hierarchy. Each concept in the hierarchy is depicted as hundreds of thousands of images collected from the Web. We used a subset of ImageNet with 1,000 concepts which were used for ILSVRC 2013.[1]

We construct a 3-layers concept hierarchy, where the 1,000 concepts are modeled as the leaf nodes. We remove the isolated nodes (the depth is less than 3). The final hierarchy consists of 1,137 concepts, including 55 concepts in the top layer (sup-categories), 180 concepts in the middle layer (mid-categories), and 902 concepts in the lowest layer (sub-categories). Among the hierarchy, two examples of the tag hierarchy of the dataset are shown in Fig. 5. This corresponds to a total of 1,106,500 training images, 45,100 validation images, and 50,000 test images. For each image instance, there are three tags (one for each layer) and they constitute a tree with a single path from the top layer to the bottom layer.

B. Experimental Setting

To test the effectiveness of our LTDL model for generating missing tags, we conduct experiments based on the constructed tag hierarchy. In order to improve the performance of hierarchical image annotation, we make our model based on the features learned from the convolutional neural network [56]. In our experiments, Deep Convolutional Activation Features of [57] are used as input, where the center only option is selected.

For the parameter setting, we set $L = 5$ and $K = 3$. The corruption level for the lowest layer is 0.1, for the middle 0.2 and for the top layer 0.3. In the training stage, the stochastic gradient decent method is used in the supervised fine-tuning stage, and as in the classic setting, we set the learning rate as 0.1.

For comparison, we compare our model with both global "flat" methods and hierarchical methods. For comparison with global "flat" methods, deep auto-encoder is used. Since there are 3 (layer 3, layer 4, and layer 5) layers in the feature hierarchies in our network, to be fair, we compare our model with the same deep non-linear modulars (Auto-encoder) and different layers, and the tag hierarchy is not used, including Stacked Denoising Auto-encoder (SDA) with 3 layers (SDA-3), 4 layers (SDA-4), and 5 layers (SDA-5). In the global "flat" methods, we predict the probabilities of the whole tags (three layers) simultaneously. Also, we compare our proposed LTDL model with two typical kinds of hierarchical methods: top-down approaches and bottom-up approaches [58]. For the top-down approach, we compare with the method analogous to [59]. In the experiment, we recursively split the set of possible labels according to the tag hierarchy and logistic regression (H-LR) is used. For the hierarchical bottom-up (H-BU) approach, we use deep auto-encoder to predict the probabilities of the tags in the bottom layer (sub-categories), and then infer the mid-categories and sup-categories based on the tag hierarchy, where the probability of the node i in the middle and top layer is the sum of probabilities of its children.

Note that there are many methods for image annotation. Especially, [56] proposed to use deep convolutional neural networks for image classification and achieved a good result. However, in this paper, we do not conduct comparative experiments with them due to the following reasons: 1) these models deal with conventional image classification instead of hierarchical image annotation, thus the problem to be solved is

[1][Online]. Available: http://www.image-net.org/challenges/LSVRC/2013/index

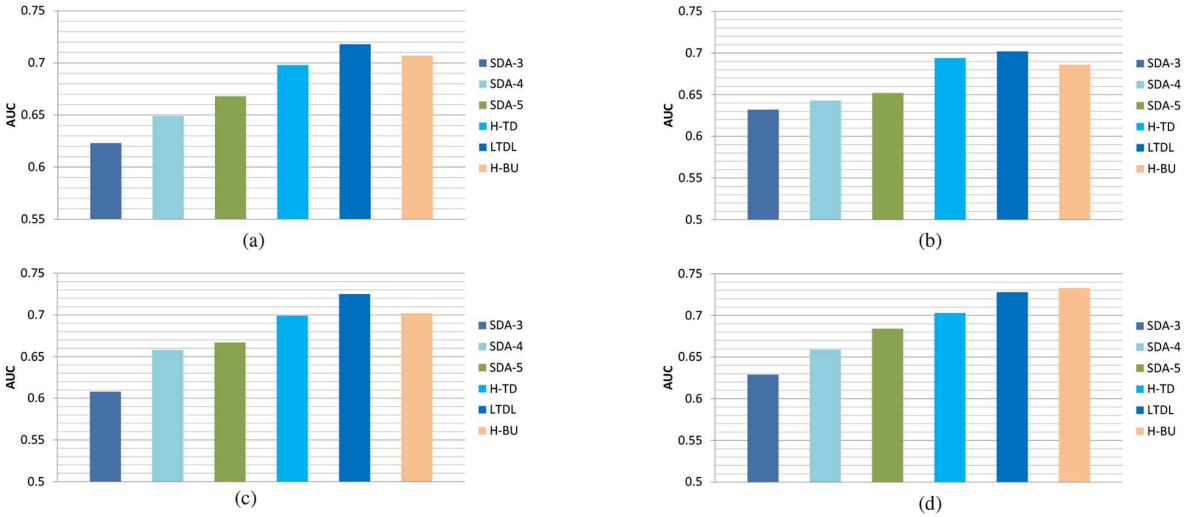


Fig. 6. AUC results for the whole categories and the categories in three layers, respectively. (a) AUC scores of the whole categories. (b) AUC scores of sub-categories. (c) AUC scores of mid-categories. (d) AUC scores of sup-categories.

different from ours; 2) we focus on investigating the usefulness of tag hierarchy information for hierarchical feature learning and our model is built on the SDA (i.e., SDA + tag hierarchy), thus we think that the comparison should be conducted with SDA, rather than CNN [56]. The effects of the network architecture cannot be eliminated if we compare with a model with a different architecture.

C. Mutli-Class AUC Evaluation

The quality of a hierarchical image annotation will be evaluated with the metric of multi-class AUC, which is directly based on the prediction probabilities. In our experiments, we show the multi-class AUC for each layer as well as the average results for the whole categories.

For each layer, we first compute the two-class $AUC(i,j)$ for the pair of tags i and j, which is computed as follows: for all the examples with class labels i and j, the prediction results are ranked in increasing sequences according to the prediction probabilities p_i (take the sample with tag i as positive sample, and take that with tag j as negative sample), and then $AUC(i,j)$ is calculated by the following equation:

$$AUC(i,j) = \frac{S_i - n_i\left(n_i+1\right)/2}{n_i n_j} \tag{8}$$

where n_i and n_j are the number of positive and negative samples respectively, and $S_i = \sum Pos(q)$, where $Pos(q)$ is the rank position of the q-th positive example in the ranked list. The higher the positive examples are ranked (with higher probability of being a positive example), the higher the term S_i will be. Therefore, AUC measures the quality of ranking, which is a more elegant metric in our problem than accuracy.

Multi-class AUC [60] is a muti-class generalization of the two-class AUC, which calculates the overall performance by averaging the two-class AUC over all pairs of classes

$$AUC_{multi-class} = \frac{2}{|c|\left(|c|-1\right)} \sum_{i<j} \widehat{AUC}(i,j) \tag{9}$$

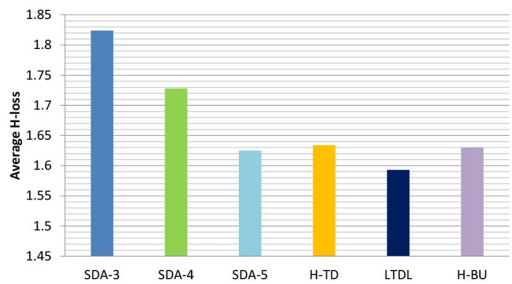


Fig. 7. Average hierarchical loss for different models.

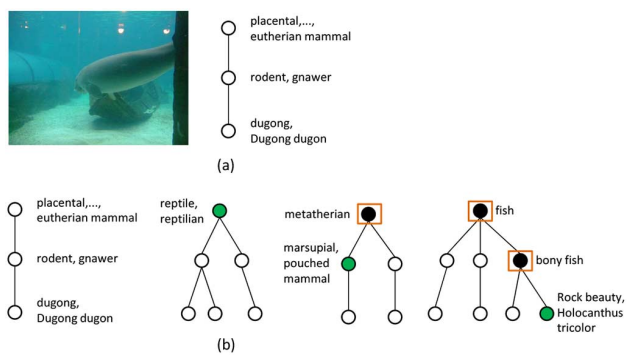


Fig. 8. Example image and associated ground-truth tags (a) and the annotation results from LTDL model (b).

where $|c|$ is the number of classes in the current layer, and $\widehat{AUC}(i,j) = \frac{AUC(i,j)+AUC(j,i)}{2}$, which is the measure of separability between classes i and j.

The experimental results of the whole average AUC and the separate AUC scores for each layer in the tag hierarchy are shown in Fig. 6. Although the AUC score of the H-BU method is slightly better than our method for the sup-categories, on the whole, LTDL outperforms other methods. From the experiments, we can draw the following conclusions: 1) Compared with the global "flat" methods without using the tag information,

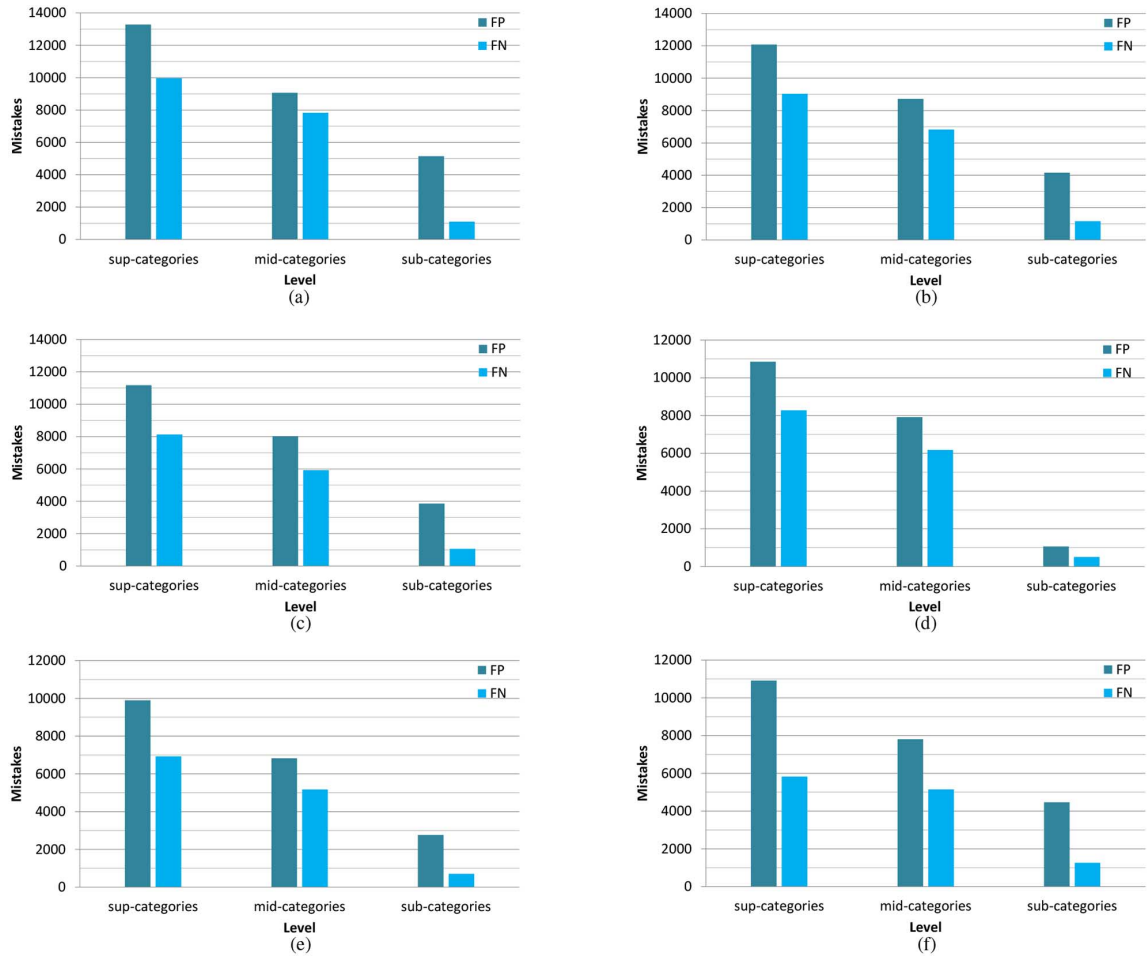


Fig. 9. Plot of the average values contained for the H-loss mistake distribution over hierarchy levels. (a) SDA-3. (b) SDA-4. (c) SDA-5. (d) H-TD. (e) LTDL. (f) H-BU.

our model integrating the tag information into the visual content learns more representative features, which is reflected in the higher AUC scores. The tags are helpful for improving the feature learning, and the layer-wise fusion mechanism in our model is effective; 2) The multi-class AUC is based on the probabilities of the tags in each layer and not dependent on the taxonomy information. However, in the H-LR and H-BU methods, the lexical taxonomy of tag hierarchy is utilized in the process of generating prediction probabilities, while in our LTDL model, the probabilities are generated by the learned hierarchical features absolutely, which demonstrates the more representative ability of our learned feature hierarchies.

D. Hierarchical Loss Evaluation

We also evaluate the methods with the metric of hierarchical loss [53] that considers the taxonomical structure to measure the discrepancy between the predicted multilabel $\widehat{\mathbf{y}} = (\widehat{y}_i, \ldots, \widehat{y}_M)$ and the true multilabel $\mathbf{y} = (y_i, \ldots, y_M)$. The leading idea underlying our hierarchical loss function is: if a parent class has been predicted wrongly, then errors in the children should not be taken into account. The loss is defined as

$$l_H(\widehat{\mathbf{y}}, \mathbf{y}) = \sum_{i=1}^{M} c_i \{\widehat{y}_i \neq y_i \wedge \widehat{y}_j = y_j, j \in ANC(i)\} \quad (10)$$

where $c_i, \ldots, c_M > 0$ are fixed cost efficients, which are set as 1 (uniform H-loss) in our experiments, and $ANC(i)$ denotes the set of ancestors of i.

The average hierarchical loss for different models can be found in Fig. 7. From the results, we can see that our LTDL outperforms the other methods. Despite all this, we find that sometimes LTDL fails to deliver correct results. Especially, compared with hierarchical methods, LTDL may give the annotation results that the predicted tag in sup-level is semantically inconsistent with that in mid-level, and the tag in mid-level is semantically inconsistent that in sub-level. Fig. 8 shows an image and associated ground-truth tags (a), and the annotation results from our LTDL model (b), where the other irrelevant tags are omitted. The tags filled with green are the predicted tags directly from the corresponding learned features, and the tags with an orange box are annotated tags through multipath labeling and partial-labeling step. In this case, the LTDL predicts the wrong tags for each layer, and the tags "reptile, reptilian", "marsupial, pouched mammal", "Rock beauty, Holocanthus tricolor" are not in the same trees. This not only leads to high H-loss (4 in this case), but means that the semantics of the predicted tags in each layer are very different. We think that this phenomenon is caused by the missing prior on tag relationships. Although the layer-wise supervised training and

multimodal fusion are used in the model, the finer granular relationships between tags are not embedded.

In order to get insight into the distribution of mistakes across the different hierarchy layers, we also evaluate the H-loss mistakes made at each layer as [53]. The nodes i makes an H-loss mistake if

$$\widehat{y}_i \neq y_i \wedge \widehat{y}_j = y_j = 1, j \in ANC(i). \tag{11}$$

Further on, we consider the false positive (FP) and false negative (FN) mistakes. Node i makes a false positive if

$$\widehat{y}_i = 1 \wedge y_i = 0 \wedge \widehat{y}_j = y_j = 1, j \in ANC(i) \tag{12}$$

and makes a false negative if

$$\widehat{y}_i = 0 \wedge y_i = 1 \wedge \widehat{y}_j = y_j = 1, j \in ANC(i). \tag{13}$$

Fig. 9 shows the distribution across the hierarchical layers of the two kinds of mistakes in our experiments. From the results, we can see that our LTDL model makes fewer mistakes and achieves better performance. Besides, we find that the number of false positive mistakes is bigger than that of the false negative mistakes, which results from the fact that our predicted multi-label is multi-path trees or partial-path trees while the true multi-label is a single-path tree.

V. Conclusion

In this paper, we propose a novel idea to explore the intrinsic hierarchical structure for images from the perspective of feature learning. Different from the current trend in multimedia analysis of using pre-defined features or focusing on the end-task "flat" representation, we propose a novel Layer-wise Tag-embedded Deep Learning (LTDL) model to learn the hierarchical features which correspond to hierarchical semantic structures in tag hierarchies. In the model, the supervised training and the multimodal fusion alternate in a layer-wise way.

Based on the LTDL model, we develop a hierarchical image annotation system, which generates the tags in a hierarchical way. We evaluate the effectiveness of the model on the public large-scale ImageNet dataset. The experimental results have demonstrated the effectiveness of the LTDL model.

For the future work, firstly, we will investigate the methods that make the property of feature hierarchy more explicit with better feature interpretation. Secondly, since the tags in the tag hierarchies have not only the inter-layer relationship, but also the intra-layer relationship, we will improve our methods so as to address the relationship between tags in the same layer. Thirdly, the idea and model should be evaluated in more social media tasks and applications, and we will investigate in particular whether the derived feature hierarchies are helpful to address other multimedia applications, such as social image management, structural image retrieval, etc.

References

[1] Y. Bengio, A. Courville, and P. Vincent, "Representation learning: A review and new perspectives," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 8, pp. 1798–1828, Aug. 2013.

[2] Q. He, K. Chang, E.-P. Lim, and J. Zhang, "Bursty feature representation for clustering text streams," in *Proc. SIAM SDM*, 2007, pp. 491–496.

[3] T. Evgeniou, M. Pontil, C. Papageorgiou, and T. Poggio, "Image representations and feature selection for multimedia database search," *IEEE Trans. Knowl. Data Eng.*, vol. 15, no. 4, pp. 911–920, Jul.–Aug. 2003.

[4] D. G. lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 91–110, 2004.

[5] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recog.*, Jun. 2005, vol. 1, pp. 886–893.

[6] H. Hotelling, "Analysis of a complex of statistical variables into principal components," *J. Edu. Psychol.*, vol. 24, no. 6, p. 417, 1933.

[7] P. Comon, "Independent component analysis, a new concept?," *Signal Process.*, vol. 36, no. 3, pp. 287–314, 1994.

[8] H. Lee, A. Battle, R. Raina, and A. Y. Ng, "Efficient sparse coding algorithms," *Adv. Neural Inf. Process. Syst.*, vol. 19, pp. 801–808, 2006.

[9] M. M. Botvinick, "Hierarchical models of behavior and prefrontal function," *Trends Cognitive Sci.*, vol. 12, no. 5, pp. 201–208, 2008.

[10] D. J. Felleman and D. C. Van Essen, "Distributed hierarchical processing in the primate cerebral cortex," *Cerebral Cortex*, vol. 1, no. 1, pp. 1–47, 1991.

[11] J. B. Levitt, D. C. Kiper, and J. A. Movshon, "Receptive fields and functional architecture of macaque V2," *J. Neurophysiol.*, vol. 71, no. 6, pp. 2517–2542, 1994.

[12] S. Osindero, M. Welling, and G. E. Hinton, "Topographic product models applied to natural scene statistics," *Neural Comput.*, vol. 18, no. 2, pp. 381–414, 2006.

[13] Y. Bengio and Y. LeCun, "Scaling learning algorithms towards AI," *Large-Scale Kernel Mach.*, vol. 34, pp. 1–41, 2007.

[14] M. Ranzato, F. J. Huang, Y.-L. Boureau, and Y. Lecun, "Unsupervised learning of invariant feature hierarchies with applications to object recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2007, pp. 1–8.

[15] H. Schulz and S. Behnke, "Deep learning: Layer-wise learning of feature hierarchies," *Künstliche Intell.*, vol. 26, no. 4, pp. 357–363, 2012.

[16] N. Srivastava and R. Salakhutdinov, "Multimodal learning with deep Boltzmann machines," *Adv. Neural Inf. Process. Syst.*, pp. 2231–2239, 2012.

[17] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Y. Ng, "Multimodal deep learning," in *Proc. 28th Int. Conf. Mach. Learn.*, 2011, pp. 689–696.

[18] R. Poli, M. Healy, and A. Kameas, *Theory and Applications of Ontology: Computer Applications*. New York, NY, USA: Springer, 2010.

[19] A. Coates, A. Y. Ng, and H. Lee, "An analysis of single-layer networks in unsupervised feature learning," in *Proc. Int. Conf. AI Statist.*, 2011, pp. 215–223.

[20] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet allocation," *J. Mach. Learn. Res.*, vol. 3, pp. 993–1022, 2003.

[21] N. Zhou, Y. Shen, J. Peng, and J. Fan, "Learning inter-related visual dictionary for object recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2012, pp. 3490–3497.

[22] N. Zhou and J. Fan, "Jointly learning visually correlated dictionaries for large-scale visual recognition applications," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 4, pp. 715–730, Apr. 2013.

[23] L. Shen, S. Wang, G. Sun, S. Jiang, and Q. Huang, "Multi-level discriminative dictionary learning towards hierarchical visual categorization," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2013, pp. 383–390.

[24] H. Zhang, Y. Zhuang, and F. Wu, "Cross-modal correlation learning for clustering on image-audio dataset," in *Proc. 15th Int. Conf. Multimedia*, 2007, pp. 273–276.

[25] K. Wang, R. He, W. Wang, L. Wang, and T. Tan, "Learning coupled feature spaces for cross-modal matching," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 2088–2095.

[26] M. Katsurai, T. Ogawa, and M. Haseyama, "A cross-modal approach for extracting semantic relationships between concepts using tagged images," *IEEE Trans. Multimedia*, vol. 16, no. 4, pp. 1059–1074, Jun. 2014.

[27] Y. Bengio, "Learning deep architectures for AI," *Found. Trends Mach. Learn.*, vol. 2, no. 1, pp. 1–127, 2009.

[28] G. E. Hinton and R. R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," *Science*, vol. 313, pp. 504–507, 2006.

[29] G. E. Hinton, S. Osindero, and Y.-W. Teh, "A fast learning algorithm for deep belief nets," *Neural Comput.*, vol. 18, no. 7, pp. 1527–1554, 2006.

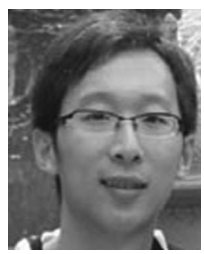
[30] P. Vincent, H. Larochelle, Y. Bengio, and P.-A. Manzagol, "Extracting and composing robust features with denoising autoencoders," in *Proc. 25th Int. Conf. on Mach. Learn.*, 2008, pp. 1096–1103.
[31] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proc. IEEE*, vol. 86, no. 11, pp. 2278–2324, Nov. 1998.
[32] M. Naphade, J. R. Smith, J. Tesic, S.-F. Chang, W. Hsu, L. Kennedy, A. Hauptmann, and J. Curtis, "Large-scale concept ontology for multimedia," *IEEE Multimedia*, vol. 13, no. 3, pp. 86–91, Jul.–Sep. 2006.
[33] M. Marszalek and C. Schmid, "Semantic hierarchies for visual object recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2007, pp. 1–7.
[34] M. Koskela, A. F. Smeaton, and J. Laaksonen, "Measuring concept similarities in multimedia ontologies: Analysis and evaluations," *IEEE Trans. Multimedia*, vol. 9, no. 5, pp. 912–922, Aug. 2007.
[35] T. Deselaers and V. Ferrari, "Visual and semantic similarity in ImageNet," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2011, pp. 1777–1784.
[36] J. Deng, A. C. Berg, and L. Fei-Fei, "Hierarchical semantic indexing for large scale image retrieval," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2011, pp. 785–792.
[37] A. Hotho, A. Maedche, and S. Staab, "Ontology-based text document clustering," *KI*, vol. 16, no. 4, pp. 48–54, 2002.
[38] A. Gelman and J. Hill, *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge, U.K.: Cambridge Univ. Press, 2007.
[39] D. M. Blei, T. L. Griffiths, M. I. Jordan, and J. B. Tenenbaum, "Hierarchical topic models and the nested Chinese restaurant process," *Adv. Neural Inf. Process. Syst.*, vol. 16, pp. 17–24, 2003.
[40] G. Irie, D. Liu, Z. Li, and S.-F. Chang, "A Bayesian approach to multimodal visual dictionary learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2013, pp. 329–336.
[41] J. Sang and C. Xu, "Faceted subtopic retrieval: Exploiting the topic hierarchy via a multi-modal framework," *J. Multimedia*, vol. 7, no. 1, pp. 9–20, 2012.
[42] J. Sang and C. Xu, "Browse by chunks: Topic mining and organizing on web-scale social media," *ACM Trans. Multimedia Comput., Commun., Appl.*, vol. 7, no. 1, p. 30, 2011.
[43] Q. Huang, M. Han, B. Wu, and S. Ioffe, "A hierarchical conditional random field model for labeling and segmenting images of street scenes," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2011, pp. 1953–1960.
[44] G. Roig, X. Boix, F. De la Torre, J. Serrat, and C. Vilella, "Hierarchical CRF with product label spaces for parts-based models," in *Proc. IEEE Int. Conf. Automat. Face Gesture Recog. Workshops*, Mar. 2011, pp. 657–664.
[45] L. Ladicky, C. Russell, P. Kohli, and P. H. Torr, "Associative hierarchical CRFs for object class image segmentation," in *Proc. IEEE 12th Int. Conf. Comput. Vis.*, Sep.–Oct. 2009, pp. 739–746.
[46] C. Collet, P. Thourel, P. Pérez, and P. Bouthemy, "Hierarchical MRF modeling for sonar picture segmentation," in *Proc. Int. Conf. Image Process.*, Sep. 1996, vol. 3, pp. 979–982.
[47] Y. Chen and T. S. Huang, "Hierarchical MRF model for model-based multi-object tracking," in *Proc. Int. Conf. Image Process.*, Oct. 2001, vol. 1, pp. 385–388.
[48] K. Kavukcuoglu, P. Sermanet, Y.-L. Boureau, K. Gregor, M. Mathieu, and Y. L. Cun, "Learning convolutional feature hierarchies for visual recognition," *Adv. Neural Inf. Process. Syst.*, vol. 1, pp. 1090–1098, 2010.
[49] E. Hörster and R. Lienhart, "Deep networks for image retrieval on large-scale databases," in *Proc. 16th ACM Int. Conf. Multimedia*, 2008, pp. 643–646.
[50] N. Cesa-Bianchi, C. Gentile, and L. Zaniboni, "Hierarchical classification: Combining Bayes with SVM," in *Proc. 23rd Int. Conf. Mach. Learn.*, 2006, pp. 177–184.
[51] G. E. Hinton, "Learning multiple layers of representation," *Trends Cognitive Sci.*, vol. 11, no. 10, pp. 428–434, 2007.
[52] Y. Bengio, P. Lamblin, D. Popovici, and H. Larochelle, "Greedy layer-wise training of deep networks," *Adv. Neural Inf. Process. Syst.*, vol. 19, pp. 153–160, 2007.
[53] N. Cesa-Bianchi, C. Gentile, and L. Zaniboni, "Incremental algorithms for hierarchical classification," *J. Mach. Learn. Res.*, vol. 7, pp. 31–54, 2006.
[54] T. Fawcett, "Roc graphs: Notes and practical considerations for researchers," *Mach. Learn.*, vol. 31, pp. 1–38, 2004.
[55] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2009, pp. 248–255.
[56] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Adv. Neural Inf. Process. Syst.*, vol. 25, pp. 1097–1105, 2012.
[57] J. Donahue, Y. Jia, O. Vinyals, J. Hoffman, N. Zhang, E. Tzeng, and T. Darrell, Cornell Univ. Library, Ithaca, NY, USA, "Decaf: A deep convolutional activation feature for generic visual recognition," *arXiv preprint arXiv:1310.1531*, 2013.
[58] C. N. Silla, Jr and A. A. Freitas, "A survey of hierarchical classification across different application domains," *Data Mining Knowl. Discovery*, vol. 22, no. 1–2, pp. 31–72, 2011.
[59] G. Griffin and P. Perona, "Learning and using taxonomies for fast visual categorization," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2008, pp. 1–8.
[60] D. J. Hand and R. J. Till, "A simple generalisation of the area under the roc curve for multiple class classification problems," *Mach. Learn.*, vol. 45, no. 2, pp. 171–186, 2001.

Zhaoquan Yuan received the B.E. degree in computer science and technology from the University of Science and Technology of China, Hefei, China, and he is currently working toward the Ph.D. degree at the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing, China.

His current research interests include social media mining, machine learning, and pattern recognition.

Changsheng Xu (M'97–SM'99–F'14) is a Professor with the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing, China, and an Executive Director with the China-Singapore Institute of Digital Media, Singapore. He holds 30 granted/pending patents and has authored or coauthored over 200 refereed research papers. His current research interests include multimedia content analysis/indexing/retrieval, pattern recognition, and computer vision.

Dr. Xu is an ACM Distinguished Scientist. He is an Associate Editor of the IEEE TRANSACTIONS ON MULTIMEDIA and *ACM Transactions on Multimedia Computing and Communications*. He served as a Program Chair of ACM Multimedia in 2009. He has served as an Associate Editor, Guest Editor, General Chair, Program Chair, Area/Track Chair, Special Session Organizer, Session Chair and TPC Member for over 20 prestigious IEEE and ACM multimedia journals, conferences, and workshops.



Jitao Sang received the B.E. degree from SouthEast University, Nanjing, China, in 2007, and the Ph.D. degree from the Chinese Academy of Sciences, Beijing, China, in 2012.

He is an Assistant Professor with the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing, China. His research interests include multimedia content analysis, social media mining, and social network analysis.

Dr. Sang has served as Guest Editor and Organizing Committee Member for several journals and conferences. He was awarded the Special Prize of President Scholarship by the Chinese Academy of Sciences. He coauthored the Best Student Paper in Internet Multimedia Modeling 2013 and the Best Paper Finalist in ACM Multimedia 2012 and 2013.

Shuicheng Yan (M'06–SM'09) is currently an Associate Professor with the Department of Electrical and Computer Engineering, National University of Singapore, Singapore, and the Founding Lead of the Learning and Vision Research Group, Singapore. He has authored or coauthored nearly 300 technical papers over a wide range of research topics. His research areas include computer vision, multimedia and machine learning.

Dr. Yan is an Associate Editor of the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY and the *ACM Transactions on Intelligent Systems and Technology*, and has been serving as the Guest Editor of the special issues for the IEEE TRANSACTIONS ON MULTIMEDIA and *CVIU*. He received the Best Paper Awards from PCM '11, ACM MM '10, ICME '10, and ICIMCS '09, the Winner Prize of the classification task in both PASCAL VOC '10 and PASCAL VOC '11, the Honorable Mention Prize of the detection task in PASCAL VOC '10, 2010 TCSVT Best Associate Editor Award, 2010 Young Faculty Research Award, 2011 Singapore Young Scientist Award, 2012 NUS Young Researcher Award, and was a coauthor of the Best Student Paper Awards of PREMIA '09, PREMIA '11, and PREMIA '12.

M. Shamim Hossain (S'03–M'07–SM'09) received the Ph.D. degree in electrical and computer engineering from the University of Ottawa, Ottawa, ON, Canada.

He is an Associate Professor with King Saud University, Riyadh, Saudi Arabia. He has authored or coauthored more than 70 publications including refereed IEEE/ACM/Springer/Elsevier journals, conference papers, books, and book chapters. His research interests include serious games, cloud and multimedia for health care, resource provisioning for big data processing on media clouds, and biologically-inspired approaches for multimedia and software system.

Dr. Hossain is a member of ACM and ACM SIGMM. He has served as a member of the organizing and technical committees of several international conferences and workshops. He served as a Co-Chair of the 1st, 2nd, 3rd, 4th, and 5th IEEE ICME Workshop on Multimedia Services and Tools for E-Health. He served as a Co-Chair of the 1st Cloud-Based Multimedia Services and Tools for E-Health Workshop 2012 with ACM Multimedia. He currently serves as a Co-Chair of the 4th IEEE ICME Workshop on Multimedia Services and Tools for E-Health. He is on the editorial board of the *Springer International Journal of Multimedia Tools and Applications*. He was on the editorial board of some journals including the *International Journal of Advanced Media and Communication*. Previously, he served as a Guest Editor of the IEEE TRANSACTIONS ON INFORMATION TECHNOLOGY IN BIOMEDICINE and *Springer Multimedia Tools and Applications*. Currently, he serves as a Lead Guest Editor of the *Elsevier Future Generation Computer Systems*, *International Journal of Distributed Sensor Networks*, and *Springer Cluster Computing*.