

Accepted Manuscript

A Region-based Image Caption Generator with Refined Descriptions

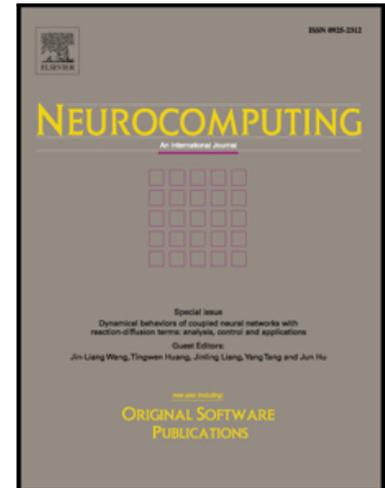
Philip Kinghorn , Li Zhang , Ling Shao

PII: S0925-2312(17)31251-1
DOI: [10.1016/j.neucom.2017.07.014](https://doi.org/10.1016/j.neucom.2017.07.014)
Reference: NEUCOM 18702

To appear in: *Neurocomputing*

Received date: 26 December 2016
Revised date: 8 April 2017
Accepted date: 6 July 2017

Please cite this article as: Philip Kinghorn , Li Zhang , Ling Shao , A Region-based Image Caption Generator with Refined Descriptions, *Neurocomputing* (2017), doi: [10.1016/j.neucom.2017.07.014](https://doi.org/10.1016/j.neucom.2017.07.014)



This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

A Region-based Image Caption Generator with Refined Descriptions

Philip Kinghorn¹, Li Zhang¹ and Ling Shao²

¹Computational Intelligence Research Group
Department of Computing Science and Digital Technologies
Faculty of Engineering and Environment
University of Northumbria
Newcastle, NE1 8ST, UK

²School of Computing Sciences
University of East Anglia
Norwich, UK

Email: {philip.kinghorn; li.zhang}@northumbria.ac.uk; ling.shao@ieee.org

Abstract.

Describing the content of an image is a challenging task. To enable detailed description, it requires the detection and recognition of objects, people, relationships and associated attributes. Currently, the majority of the existing research relies on holistic techniques, which may lose details relating to important aspects in a scene. In order to deal with such a challenge, we propose a novel region-based deep learning architecture for image description generation. It employs a regional object detector, recurrent neural network (RNN)-based attribute prediction, and an encoder-decoder language generator embedded with two RNNs to produce refined and detailed descriptions of a given image. Most importantly, the proposed system focuses on a local based approach to further improve upon existing holistic methods, which relates specifically to image regions of people and objects in an image. Evaluated with the IAPR TC-12 dataset, the proposed system shows impressive performance, and outperforms state-of-the-art methods using various evaluation metrics. In particular, the proposed system shows superiority over existing methods when dealing with cross-domain indoor scene images.

Keywords: Image Description Generation, Convolutional and Recurrent Neural Networks, Description Generation.

1 INTRODUCTION

Describing the contents of images is a relatively easy task for humans. Humans can recognize, distinguish, and describe objects, scenes or actions influenced by many external factors, such as occlusions, illumination changes and pose variations, which make this task incredibly difficult for machines to compete [1]. Recent research shows that state-of-the-art computer vision algorithms have become incredibly powerful in domains such as image classification and segmentation. However, generation of refined descriptions of image content is still a difficult task.

In recent years, many methods have been proposed for image description generation. However, the majority of the related research relies on holistic approaches for image understanding and entity recognition, which may lose details relating to important aspects of an image. In order to achieve refined and detailed description, this research aims to propose a novel local deep learning architecture for image description generation. It employs a regional object detector, recurrent neural network (RNN)-based attribute classification, and a pair of encoder-decoder based RNNs to generate detailed descriptions of image contents. Most importantly, the proposed system focuses on a local based approach to improve upon existing holistic methods, which relates to image regions of people and objects in a given image.

The proposed system consists of four key stages: (1) object detection and recognition; (2) attribute prediction; (3) scene classification; and (4) description generation. The overall system architecture is shown in Figure 1. In the first stage, an object detector is implemented with the use of a large scale deep Convolutional Neural Network (CNN) to locate and classify people and objects within an image. This CNN provides bounding boxes and object class labels as outputs. In the second stage, the above CNN is applied to the detected regions to extract features for subsequent attribute prediction. Two RNNs are implemented for attribute classification from the local regions, with one dedicated to human attribute prediction and the other applied to object attribute prediction. In the third stage, machine learned holistic image features are extracted using the above CNN for scene classification. In the fourth stage, the recognized objects, scene, people and their associated attribute labels are passed to an encoder-decoder structure, which consists of two RNNs to translate class and attribute labels to full descriptions.

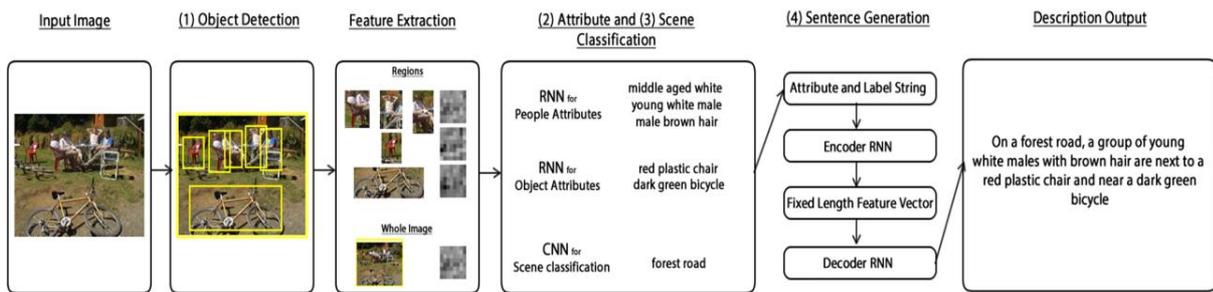


Figure 1 The overall system architecture, which consists of (1) object detection and recognition, (2) attribute prediction, (3) scene classification and (4) sentence generation

The main contributions of our research are two-fold, as follows.

- A local region-based deep learning architecture is proposed for image description generation. In order to overcome limitations of existing holistic methods, it employs a regional object detector, RNN-based attribute prediction, and encoder-decoder based description generation. Especially, the challenge of sentence-based captioning and description generation in this research is treated as a machine translation problem.
- A comprehensive evaluation of the proposed system is conducted using the IAPR TC-12 dataset. The empirical results indicate that the proposed system shows impressive performance and outperforms state-of-the-art related research on nearly all the evaluated metrics. In particular, the system shows great superiority and efficiency when dealing with cross-domain indoor scene images extracted from the NYUv2 sentence dataset, which is wholly different from the dataset used for training the system and is regarded as a great challenge to existing research.

The paper is organized as follows. Section 2 discusses state-of-the-art related research. In Section 3, we present the proposed system, including object detection and recognition, scene classification, RNN-based attribute prediction, and encoder-decoder based sentence generation. A comprehensive evaluation is provided in Section 4. Section 5 summarizes our work and identifies potential areas for future research.

2 RELATED WORK

In this section, we review state-of-the-art methods on object detection and recognition, attribute prediction and image description/caption generation.

2.1 Object Detection and Recognition

Girshick et al. [2] proposed an object detection algorithm with semantic segmentation by combining region proposals with CNN features. Known as R-CNN, their system extracted region proposals and employed the CNN for subsequent feature learning to generate a fixed dimensional feature vector for each proposal. Linear Support Vector Machines (SVMs) were used to classify each region. Their system was able to outperform the sliding window CNN for object detection, such as OverFeat [3], significantly.

2.2. Attribute Prediction

Attributes have been widely used in numerous areas of research, either by adding high level information to labels or by using them to aid other computer vision tasks such as scene or object classification [4-7]. Farhadi et al. [4] described objects by their attributes, enabling objects that could not be classified into a specific object label to be described by their shape, size, color, or texture. Their system utilized edge detection, feature extraction using HOG descriptors, and color extraction using LAB-based descriptors. Attributes were used by Dhar et al. [5] for predicting aesthetics and interestingness. Their work demonstrated predictions for compositional attributes. They also employed attributes to represent the presence of objects, animals and/or people, as well as illumination. To determine ‘interestingness’ of an image, color histograms, Harr features, and spatial pyramid, were extracted as well. Research from Bourdev et al. [6] used pose information, called poselets, to determine attributes of a person, such as gender and clothing. Their system consisted of four stages. Stage 1 detected all poselets in an image. In stage 2, feature vectors were extracted for each poselet type. In stage 3, *poselet-level attribute classifiers* were used to predict the presence of an attribute. In stage 4, a linear *person-level attribute classifier* was used to combine the evidence from all body parts for attribute prediction. Lang et al. [7] fused image features and visual attributes to conduct covert photo classification. They first processed each image to generate a spatial pyramid. Then, for each sub-image, features were extracted for visual attribute classification. Their work employed 13 visual attributes including image brightness, colour richness, etc. The fusion of the image features and the visual attributes was then used for covert image classification.

Although the above related work obtained impressive performance for attribute prediction, the majority of such applications employ a classifier cluster for attribute prediction with each individual classifier trained on a specific set of features relating to the respective attribute label [6, 7]. Such methods can only label each attribute as present or otherwise. This research aims to overcome such limitations by employing RNNs for the prediction of highly related attributes, with the aim to increase description capability of the proposed system.

2.3 Image Description Generation

Automatic image captioning has drawn great attention in recent years [8-21]. Karpathy and Li [8] proposed a system to provide natural language descriptions for image regions. It employed a CNN to extract features from image regions and an RNN to generate short sentence descriptions for each region. A version of their system, known as NeuralTalk, also enables it to generate descriptions for the whole image. The work was evaluated using the Flickr8K, Flickr30K, and MSCOCO datasets and achieved impressive performance. Vinyals et al. [9] also proposed a holistic method based neural image caption generator, known as NIC. It integrated a deep CNN as the image encoder for vision feature learning and an RNN for caption generation. Significantly higher BLEU scores were achieved in comparison with those from other state-of-the-art systems when evaluated using PASCAL, Flickr8k, Flickr30k, and SBU datasets. However, their work did not extract any high-level semantic information such as attributes or relationships for caption generation.

Lin et al. [10] proposed a framework for generating multi-sentence lingual descriptions of complex indoor scenes. Their work focused on the development of a 3D parsing system in order to generate a semantic representation of the scene. Specifically, a holistic conditional random field (CRF) was employed for the scene graph generation. Their results indicated that their CRF-based description generation [11] was susceptible to a lack of diversity, owing to simply listing all results to form an overall description. Recent work by Mathews et al. [12] attempted to address a limitation in image captioning systems by incorporating sentiment elements. The architecture of their system combined CNN-based ‘factual’ caption generation with RNN-based sentiment word generation, to produce the final description. As an example, the output could change from ‘a black and white cat lying on a bed’ to ‘a close up of an adorable cat lying on a couch’.

As a well-known image captioning framework, the work of Xu et al. [13], known as Show, Attend and Tell, incorporates an attention model within the RNN, allowing the model to learn which salient regions it should focus on for the generation of the next sequence of words. It employs the well-known structure of CNN + RNN. Their attention model can also be altered to include hard or soft attention mechanisms. Lu et al. [14] produces a framework, known as Adaptive Attention. Their work does not require the attention model to be active for the generation of each word. For instance, attention and visual information does not need to be utilized for the generation of words such as ‘and’, ‘the’, as well as other words that can be inferred from the language model without the need for visual input. Their model determines at each time step whether and where the visual attention is required. Moreover, both attention models employed in [13] and [14] attempt to mimic the way that humans look at

images, with the intention to focus on the most important aspects of an image. Liu et al. [15] proposed an ‘Image2Text’ framework, which is also based upon the conventional CNN + RNN structure. However it replaces the CNN with the detected object representations, which are then passed to the RNN for caption generation.

Proposed by Johnson et al. [16], DenseCap focuses on the production of detailed regional descriptions rather than the typical full image captions. The work also follows a similar principle on notable regions by combining a Regional Proposal Network (RPN), which is based on the work of Ren [17]. This RPN component is inserted, as a localization layer, between the CNN-based feature extraction and the RNN-based language model. DenseCap is similar in effect to the work of Mao et al. [18], in which it aims to describe specific objects or regions within an image.

Recent work of You et al. [19] aims to combine both top-down and bottom-up approaches in the field of image captioning. This is achieved by extracting CNN features and using them to give the system an initial overview understanding of the image. This then allows attention mechanisms to attend to cognitive cues within the image which are used to generate subsequent words in their RNN structure. To generate attributes, it proposed two approaches, i.e. (1) a non-parametric approach which utilizes nearest neighbor image retrieval with a large corpus of paired data, and (2) two parametric models, trained with ranking-loss and Fully Convolutional Network, respectively, for attribute prediction. These two approaches can be used in conjunction if required.

3 THE PROPOSED IMAGE DESCRIPTION GENERATION SYSTEM

As discussed earlier, state-of-the-art methods largely rely on holistic techniques to extract image features for caption generation [8, 9]. The major drawback of such methods (e.g. NIC) is that local information could be overlooked. In order to overcome such limitations, this research proposes a region-based deep learning approach to capture local details by incorporating object detection and recognition, scene classification, attribute prediction and description generation. Moreover, the captioning and description generation challenge is treated as a machine translation problem, therefore an encoder-decoder structure for sentence generation is proposed. The proposed system possesses superiority over existing state-of-the-art methods, especially when dealing with image description tasks for out-of-domain images. We introduce each major step of the proposed system in detail in the following sub-sections.

3.1 Object Detection and Recognition

For robust object detection, we implement the R-CNN [2] object detector. The R-CNN is a traditional CNN but with extra outputs that predict bounding box coordinates. The R-CNN is trained using the ILSVRC13 dataset. It is able to detect, localize, and classify 200 object categories. The R-CNN [2] uses the ImageNet database [22] for the training of the CNN adopted from Krizhevsky et al. [23].

Specifically, the R-CNN employed in this research for object detection applies a Selective Search (SS) algorithm [24] to collect regions of interest (ROIs) from images that possibly contain objects and/or people. A greedy search algorithm then groups similar regions and measures similarity between regions and their neighbors. This process is repeated until the whole image becomes a region. In the greedy search algorithm, the similarity between regions a and b is defined as:

$$(a, b) = Ssize(a, b) + Stexture(a, b), \quad (1)$$

which returns a value between [0, 1]. In Equation (1), $Ssize(a, b)$ is a proportion that a and b both occupy, which persuades small regions to merge, while $Stexture(a, b)$ is the intersection between SIFT like measurements. The regions are warped to fit the required input of 227x227 pixels.

For each region proposal, a feature vector with the size of 4096 is extracted. During test, every feature vector is passed to each of the 200 SVMs representing each object class to deliver a score for each region. Non-Maximum Suppression (NMS) is also used to dispose of a region if the region has an Intersection of Union (IoU) with a comparatively higher score than a predefined threshold.

The R-CNN implementation by Girschick et al. [2] is modified in this research to crop and store images, bounding boxes, their respective class labels, and confidence scores. The class labels are concatenated to form a list of detections. The detections and the regions are then cropped and labeled again, as an object or a person. This additional label is used to determine which attribute predictors (e.g., object or person attribute predictor) the system utilizes in the subsequent processing.

3.2 Scene Classification

Scene classification is employed in this research to further increase the descriptive nature of the proposed system. Instead of simply stating indoor or outdoor, our system produces a semantic scene label (e.g. a park or a shopping mall) obtained by scene classification, in order to provide a more refined description.

This scene classification employs the hybrid Alex-Net deep CNN [23], which is trained to detect 1183 categories, including 205 scene labels and 978 object categories. This network, however, cannot be applied as an object detector in our work. This is because the hybrid CNN network does not provide object localization, i.e., bounding boxes, which are essential to our application. But it is used to classify a total of 205 scene labels and ensure an efficient coverage for description generation. Overall, it achieves impressive accuracy with a relatively low computational cost while dramatically increasing the descriptive capability of the proposed system.

3.3 Attribute Prediction Using RNNs

Traditionally, attributes are normally classified using clusters of classifiers, with one classifier dedicated to one attribute [4, 6]. Such classifier clusters not only require more resources for training, but also tend to only indicate the presence or absence of an attribute, without any confidence measure.

In this research, we employ RNN-based attribute classification to address the above drawbacks. Research has indicated that RNNs are useful in many areas of machine learning, including image captioning and machine translation [9, 25]. RNNs are used in this work because of their ability to not only classify an arbitrary number of human and object attributes, but also effectively determine which attributes should be reported for a given set of features dynamically. In this research, two RNNs are used for attribute classification, with one dedicated to human attribute prediction and the other applied to object attribute classification. Splitting the people and object attribute classification aims to reduce the amount of training data required, as well as preventing human attributes from being generated for the description of an object, and vice versa.

Also, RNNs have been used for attribute prediction owing to their significance and flexibility in dealing with natural language processing tasks. The initial comparison between RNNs and other alternative methods also indicates that RNNs show better accuracy than those of other methods such as SVMs for attribute prediction.

Moreover, the RNNs employed in this work are ‘word based’ and adopt a Long Short Term Memory (LSTM) architecture [26]. At test time, the above-mentioned CNN is used to extract image features. The extracted features are then used to predict multiple attributes one by one. The attribute prediction is based on the extracted image features combined with the previously generated words. This process continues to generate relevant attributes until the designated STOP word has been generated. I.e. the generation of this STOP word occurs when the RNN determines that no other attributes could be used to describe an image, based on the features and all of the previously generated attributes.

In this research, PubFig [27] and a subset of ImageNet [22] are also used for training of RNN-based human and object attribute prediction, respectively. A slightly modified version of AlexNet from Krizhevsky et al. [23] without classification layers is implemented to extract CNN features for training of RNNs. This network extracts 4096 image features from previously cropped images of the desired objects or people, which are then paired with the attribute labels from the respective attribute dataset for training.

3.3.1 Human attributes

The RNN for human attribute prediction in this research is trained with the PubFig [27] dataset, which consists of ~10,000 images in total. It has 200 unique celebrity faces, each labeled with 73 attributes, such as age, gender, and hair style. The attributes used to describe people in this research are a selected subset of those used in the PubFig dataset. I.e. overall 26 human attributes are chosen for this work, which are shown in Table 1.

3.3.2 Object attributes

The ImageNet [22] dataset is widely used in many computer vision challenges. It also has a small subset of images that is fully annotated with object bounding boxes and their respective attributes. This subset consists of around 10,000 images collected from 400 synsets. Each synset represents a group of ImageNet images associated with a

specific WordNet [28] ID. Each image is paired with 25 object attributes, and all these 25 attributes are employed in this research. They are illustrated in Table 2.

Overall, in comparison with traditional SVM-based attribute classification, the RNN-based object and human attribute prediction shows great efficiency for the classification of highly associated attributes, and provides efficient flexibility and diversity to aid subsequent sentence generation.

Table 1 Human attributes employed in this research

| | |
|--------------------|--|
| Hair Color | brown, blonde, black, grey, etc. |
| Hair Style | wavy, curly, straight, etc. |
| Age | child, youth, middle aged, senior |
| Gender | male, female |
| Ethnic | Asian, White, Indian |
| Accessories | glasses, sunglasses, lipstick, necklace, necktie |
| Facial Hair | goatee, moustache |

Table 2 Object attributes employed in this research

| | |
|----------------|---|
| Color | black, blue, brown, grey, green, orange, pink, red, violet, white, yellow |
| Pattern | spotted, striped |
| Shape | long, round, rectangular, square |
| Texture | furry, smooth, rough, shiny, metallic, vegetation, wooden, wet |

3.4 Sentence Generation

In this research, we regard language generation as a machine translation problem. The generated sentences are intended to be more descriptive than those generated by other existing research, with attribute and labeling details owing to the proposed local based approach. To this end, the IAPR TC-12 dataset [29] is used for both training and evaluation of the proposed system owing to its detailed captions in comparison to those provided by other popular databases (such as Flickr).

In this research, an encoder-decoder RNN structure is proposed, to overcome typical challenges in the machine learning field, such as variable input length. It consists of two RNNs, with one serving as an encoder and the other as a decoder, to overcome the input length variation problem. This encoder-decoder architecture was originally proposed by Bahdanau et al. [25] for machine translation problems. We transform it to deal with sentence generation and use the encoder RNN to encode the noun (objects/people) and adjective (attributes) keywords into a fixed-length vector. The decoder RNN is employed to transform this fixed-length vector into a full descriptive sentence. The architecture of the encoder-decoder language generator is illustrated in Figure 2.

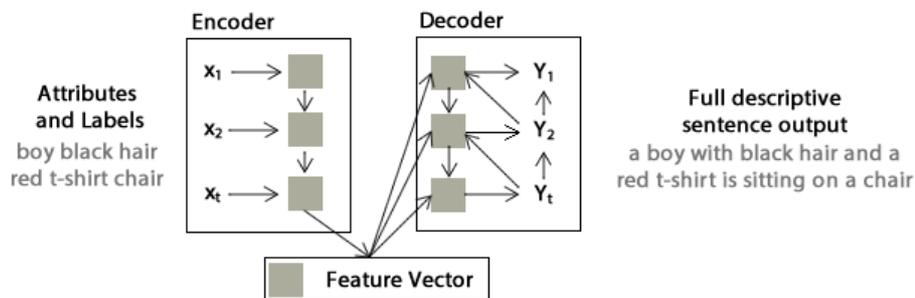


Figure 2 The encoder-decoder architecture with example attribute and label inputs and the expected sentence generation output

The proposed language generation component is trained using caption data from MSCOCO [30] and a small subset of IAPR TC-12 [29], in order to infer varying types of sentence structures. Specifically, each image in the MSCOCO [30] dataset is paired with five captions. The IAPR TC-12 dataset only has one sentence per image, however, it is more descriptive than captions provided by other datasets (e.g. Flickr30k [31]). Our proposed sentence generation structure (i.e., encoder-decoder) is therefore trained on nouns and adjectives from the captions provided by the above

training datasets. Each caption is first POS tagged with the Python NLTK (Natural Language Tool Kit) [http://www.nltk.org/], enabling nouns and adjectives to be extracted and stored. This leads to a source training set, which consists of ~30,000 unique words, and a target training set, which consists of ~20,000 unique words. The overall size of the training set is more than 500,000 captions, with more than 20,000 used as a validation set.

Table 3 Information used for sentence generation

| Recognized Entities & Attributes | Semantic Labels |
|----------------------------------|--|
| Scenes | 205 scene labels from MIT scenes including playground, living room and bedroom, etc. |
| People | Present/Non-Present |
| Objects | 200 object categories from ILSVRC13 including car, dog, orange, desk, etc. |
| Person Attributes | A subset of 26 human attribute labels taken from the PubFig dataset |
| Object Attributes | All 25 object attribute labels present in the ImageNet dataset |

Both the encoder and decoder RNNs are implemented based on [25]. All sequences longer than 50 words are ignored. Each RNN consists of 1000 hidden units in a single hidden layer to compute the probability of the next target word. These RNN models are trained with Stochastic Gradient Descent (SGD) with the use of Adadelta. During training, each update is computed with minibatches of 80 sentences. When an RNN has been fully trained, beamsearch is implemented to generate a description that maximizes the conditional probability of the source matching the target, i.e. $-argmax_y P(y/x)$. During the test stage, object/people/scene and attribute labels generated in earlier stages are used as the input source language with the full caption as the intended output. Overall, information taken into account during sentence construction is summarized in Table 3.

4 EVALUATION

To evaluate the effectiveness of the overall system for image description generation, existing research, including NeuralTalk [8], NIC [9], Show, Attend and Tell [13] and Adaptive Attention [14], has been employed for comparison. Besides that, we also evaluate the efficiency of the RNN-based attribute prediction. We first introduce the evaluation metrics applied in this research.

4.1 Evaluation Metrics

In this research, BLEU, ROUGE, and METEOR are employed for description generation evaluation. All of these methods use a similarity based measure between machine generated and ground truth sentences. We introduce each of these evaluation methods in the following sub-sections.

4.1.1 BLEU

The BLEU score [32] has been commonly employed as a machine translation measure. It can be applied to determine the semantic similarity between sentences. In this research, it is used to determine the closeness between a machine generated description and multiple high-quality human generated sentences. Moreover, in the literature, human performance has been stated to achieve a performance benchmark of 0.69 on the BLEU score [9].

4.1.2 ROUGE

ROUGE-L is a subset of all of the available ROUGE [33] metrics. It takes two input sequences into account and identifies the Longest Common Subsequence (LCS). This LCS is the subsequence that occurs in both sequences with the maximum length. In sentence level ROUGE-L, the perception is that the longer the LCS of generated and GT sentences, the more similar the sentences.

4.1.3 METEOR

The METEOR [34] score has been widely used owing to its high correlation to human subjects' annotations in comparison with other metrics [35]. It not only computes sentence similarity scores between reference and machine generated sentences, but also exhaustively identifies all matches between sentences based on certain matching criteria, such as exact word, synonym, and paraphrase matching.

4.2 Evaluation Results

To thoroughly test the proposed system, a comprehensive evaluation study is conducted. We first of all evaluate the RNN-based attribute classification using the test sets of their respective databases, i.e. the ImageNet [22] and PubFig [27] datasets. We then test the system functionality on image captioning and description generation with a substantial evaluation using the IAPR TC-12 [29] dataset, owing to its caption size and complexity. A small-scale

experiment is also conducted using the NYUv2 sentence dataset [36] to further test the system superiority and efficiency for dealing with out-of-scope images. State-of-the-art methods, such as NeuralTalk [8], NIC [9], Show, Attend and Tell [13] and Adaptive Attention [14], have also been employed for image captioning performance comparison.

4.2.1 Evaluation of attribute prediction

To compare the outputs of the RNN-based attribute classification of the proposed system with the annotations provided in the attribute datasets of PubFig and ImageNet, the BLEU score is used since multiple attributes can be treated as a sequence of words, resembling a sentence.

The PubFig dataset and the subset of ImageNet are used for evaluation of human and object attribute classification, respectively. For evaluation of object attribute prediction, the subset of ImageNet is split into ~7000 training images with ~1500 images for validation and test, respectively. Evaluated with ~1500 test images, the proposed object attribute RNN classifier achieves an impressive BLEU score of 0.62 for BLEU-1. Some example object attribute prediction results are shown in Figure 3. We also divide the PubFig dataset into ~6000, ~1000, and ~1000 images for training, validation, and test, respectively, for evaluation of human attribute classification. Based on the testing of ~1000 images, the proposed RNN achieves a BLEU score of 0.61 for BLEU-1. As observed, both scores for object and human attribute prediction are very close to the benchmark BLEU score of 0.69 pertaining to human performance [9]. Figure 4 shows some example outputs of human attribute classification.

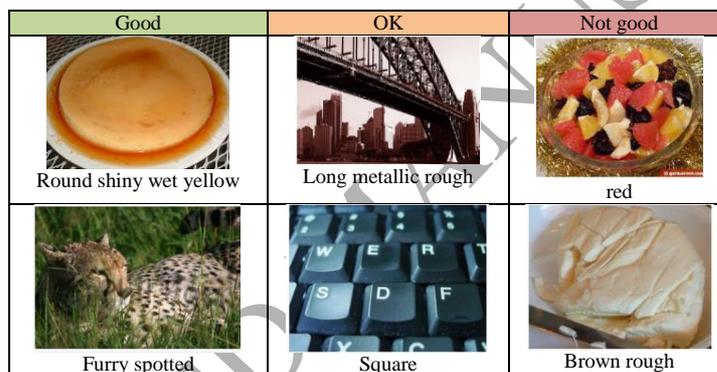


Figure 3 Object attribute classification results using object attribute RNN

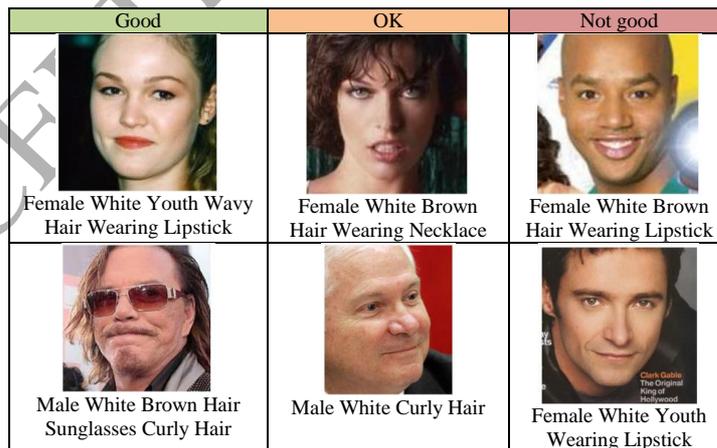


Figure 4 Person attribute classification results using human attribute RNN

4.2.2 Evaluation of image description generation

4.2.2.1 Evaluation using the IAPR TC-12 database

Evaluation of image description generation is conducted using multiple metrics, i.e., BLEU [32], ROUGE [33], and METEOR [34]. The main dataset used for evaluation is IAPR TC-12 [29], consisting of 20,000 images. This dataset is used for evaluation over existing datasets such as MSCOCO [30] or Flickr8k/30k [31], due to the characteristics of our generated descriptions. The above popular datasets such as MSCOCO and Flickr8k/30k typically consist of multiple short captions. However, the proposed system tends to produce descriptions that are more than twice the length of typical captions provided by these databases, which makes the comparison with these datasets less relevant. Moreover, the images from IAPR TC-12 are considerably larger with more objects and human actions included, paired with longer and more descriptive annotations than those in other datasets such as PASCAL [37], Flickr, and MSCOCO. Therefore it is selected in this research for evaluation.

Moreover, the proposed pipeline processing of this research also has added benefits regarding the training data required. The proposed system has been trained on a large text corpus, however the number of training images required is considerably less than those for existing methods (e.g. NIC). As mentioned earlier, we also conduct experiments with NeuralTalk [8], NIC [9], as well as the attention based methods, such as Show, Attend and Tell [13] and Adaptive Attention [14], for performance comparison. All methods are tested with their best pre-trained models. Moreover, NIC is trained on Flickr whereas the other three comparative methods are all trained on MSCOCO. However, our proposed system is trained on multiple seemingly unrelated databases (such as the ILSVRC13 dataset for object detection, the ImageNet and PubFig datasets for object and human attribute prediction, respectively). To ensure a fair comparison, an unseen test subset of 2400 images from the IAPR TC-12 dataset is used for the evaluation of all methods.

Table 4 Evaluation results for the IAPR TC-12 dataset using the MSCOCO evaluation script

| | BLEU-1 | BLEU-2 | BLEU-3 | BLEU-4 | METEOR | ROUGE-L |
|----------------------------|--------------|--------------|--------------|--------------|--------------|--------------|
| This Research | 0.231 | 0.099 | 0.046 | 0.024 | 0.067 | 0.183 |
| NeuralTalk [8] | 0.133 | 0.072 | 0.041 | 0.025 | 0.067 | 0.222 |
| NIC [9] | 0.098 | 0.048 | 0.025 | 0.015 | 0.060 | 0.209 |
| Show, Attend and Tell [13] | 0.117 | 0.067 | 0.037 | 0.021 | 0.077 | 0.204 |
| Adaptive Attention [14] | 0.052 | 0.027 | 0.014 | 0.008 | 0.060 | 0.196 |

The generated sentences by all models are compared with the reference sentences associated with each image provided by the evaluation database, and passed through the MSCOCO evaluation script [30]. This evaluation script uses all the above-mentioned metrics and enables comparison between the four baseline methods and the proposed research. Since each image in the IAPR TC-12 dataset is only paired with one description, this makes the image captioning task more challenging for all methods to score well with respect to the above metrics due to a dramatic lack of diversity.

The results for the IAPR TC-12 dataset are shown in Table 4. As illustrated in Table 4, in comparison with NeuralTalk [8], NIC [9], Show, Attend and Tell [13] and Adaptive Attention [14], our system achieves superior performance, and outperforms all other baseline methods for nearly all evaluation metrics. Specifically, in comparison with the attention based methods, i.e. Adaptive Attention [14] and Show, Attend and Tell [13], our work outperforms these two methods considerably for the BLEU metrics and remains competitive in the other metrics with Show, Attend and Tell [13] beating our score in METEOR. Moreover, it is worth to point out that all the baseline methods, especially NeuralTalk [8] and NIC [9], rely on the typical CNN + RNN structure as the foundation for caption generation, while the proposed system employs a local based approach to carry out not only local object detection and recognition, but also associated attribute prediction to inform subsequent encoder-decoder based description generation. The empirical results indicate that the sentences generated by our system are thus more descriptive. Comparatively, all baseline methods generate shorter descriptions using their respective methods. Overall, the empirical results and human inspection indicate that the generated descriptions from our system show more descriptive capabilities than those generated by NeuralTalk [8], NIC [9], Show, Attend and Tell [13] and Adaptive Attention [14]. Table 5 shows some example outputs of our system and the four baseline methods based on the evaluation of the IAPR TC-12 dataset.

Table 5 Example system outputs as compared with reference descriptions and the outputs of all baseline methods for the evaluation of the IAPR TC-12 dataset

| | | |
|-----------------------|---|---|
| IAPR TC-12 |  |  |
| Our system | View from above of a valley. | a sea cliff with a brown, red and brown cliffs and a wet seal behind it |
| Reference annotation | A river in a brown valley with many terraces; a yellowish-brown bush in the foreground. | grey and brown rocks in the foreground a light brown sandy beach at the sea and wooded hills behind it a blue sky in the background |
| NeuralTalk | a dog is running through the woods | a person on a beach with a surfboard |
| NIC | a herd of elephants walking across a lush green field | a man is standing on a rock overlooking a lake |
| Adaptive Att. | a group of people standing on top of a lush green field | a rocky beach with rocks on it |
| Show, Attend and Tell | A woman is walking down a rocky hill | A man is standing on a rock in the desert |
| IAPR TC-12 |  |  |
| Our system | three people walking across a rope bridge | a hotel room with brown walls, a rectangular bed and a wooden sofa with a black, rectangular lamp behind it |
| Reference annotation | a woman is walking on a narrow rope bridge in the middle of a forest with many green trees and bushes | a single bed made of wood with a red pillow and a striped blanket two small bedstands made of wood with a bedside lamp each a light blue carpet and a wooden wall in the background |
| NeuralTalk | a man is climbing a rock | a dog is laying on the floor with a blue ball in its mouth |
| NIC | a man is standing in the woods holding a frisbee | a bed with a white comforter and a white blanket |
| Adaptive Att. | a brown and white dog standing next to a fence | a bedroom with a bed and a dresser |
| Show, Attend and Tell | A man in a white shirt is standing in a wooded area | A woman in a pink striped shirt is sleeping on a bed |
| IAPR TC-12 |  |  |
| Our system | a construction site with a thunderous person at a round, brown wooden bicycle on a long wooden cart | a kitchen with red walls and brown vegetation on the white table |
| Reference annotation | a man with a blue cap and a greenish brown overalls is standing with a red machine in front of a grey wall a black iron door on the right | a tray on bar with cut melons oranges mangos and a coconut in form of a mouse head a man behind it three brown round doors and a white wall in the background |
| NeuralTalk | a man sitting on a bench with a dog | a woman is sitting on a couch with a man in a black shirt |
| NIC | a man in a red shirt is sitting on a bench in a park | a table with a vase of flowers and a vase |
| Adaptive Att. | a man standing next to a bicycle on a sidewalk | a woman is holding a stuffed animal in her hands |
| Show, Attend and Tell | A man in a blue shirt and blue hat is riding a bicycle | A woman in a white shirt is sitting on a table with flowers and flowers |

Referring to Table 4, it is worth mentioning that, despite the comparatively high BLEU scores and comparable METEOR scores yielded by our system, it achieves a comparatively lower ROUGE-L score than those of the other systems. This could be owing to certain characteristics of our generated descriptions. As discussed earlier, the ROUGE-L score looks for the longest common subsequence between a machine generated caption and the reference. In the proposed system, since attributes are predicted to enrich the description, it has occasionally used additional or different attributes to precede object labels. This could have a negative impact with respect to the ROUGE metric, which leads to cutting the affected sequence at that point for score calculation, therefore reducing the overall ROUGE-L result. For future work, we aim to produce a dataset similar to Flickr or MSCOCO, but with more detailed descriptive sentences to enable further evaluation of our proposed system and other similar methods.

4.2.2.2 Additional experiment using NYUv2

A small-scale additional experiment is also conducted using cross-domain images extracted from a scene description challenge, i.e. the NYUv2 sentence dataset [36]. The objective is to determine how well our system is able to cope with scene description generation for out-of-scope images. This NYUv2 sentence dataset describes purely indoor scenes with annotated descriptions containing objects, their attributes and relationships between multiple objects. Each image in this dataset is paired with one description. Each description consists of up to three sentences. Therefore, the available reference annotations for this dataset could range from very brief (e.g. one sentence) to very descriptive (e.g. multiple sentences).

In this small-scale experiment, we train the encoder-decoder based sentence generation component purely on captions and descriptions provided by the MSCOCO dataset. The newly generated model and all other baseline systems are then used to test upon ~1400 images extracted from the NYUv2 sentence dataset. Without depth information, the experiments indicate that this is a challenging dataset for the proposed system and related methods. Although the overall metric scores obtained for this dataset are comparatively lower for all models than those achieved using the IAPR TC-12 dataset, the results indicate that for the BLEU-1 metric, descriptions generated by our system outperform those generated by NeuralTalk [8], NIC [9], and Adaptive Attention [14] by 30%, 70% and 71%, respectively. It also scores equally against Show, Attend and Tell [13] under the same metric. Some example outputs generated by our system and related methods are shown in Table 6.

Table 6 Example system outputs as compared with reference descriptions and the outputs of all baseline methods for the evaluation of the NYUv2 sentence dataset

| NYUv2 Sentence |  |  |
|-----------------------|---|---|
| Our system | A hotel room with a long rectangular draped sofa and a very long, rectangular bookshelf awaits | An office setting with long rectangular red chairs next to a wooden table. |
| Reference annotation | This is a living room with wooden floor. There is a big beige sofa on the left of the room with two pillows on top. Near the sofa is a black armchair. There is a book cabinet behind the sofa and the room is separated by a door. | This is a conference room with a long wooden table with red chairs around it, a projector mounted to the ceiling, and red window blinds. A multiline phone sits on the table. |
| NeuralTalk | a man and a woman are sitting on a bench | a man is sitting on a bench in a room |
| NIC | a living room with a couch and a tv | a living room with a couch and a tv |
| Adaptive Att. | a living room filled with furniture and a bookshelf | a room with two windows and a window |
| Show, Attend and Tell | A woman sitting on a couch in a library | Two men are sitting on a couch |

4.2.3 Real-life Deployment

In order to determine how the proposed system would perform under real-life or in the wild scenarios, some initial experiments were conducted using the vision system of a humanoid NAO robot. Because of the computationally exhaustive process of the proposed system, instead of deploying it to the robot platform, we utilize a GPU server.

The robot is therefore performed as a client which communicates with our GPU server via a wireless network. This enables the robot's vision API to capture real-life images, which are subsequently transferred to and analyzed in the remote GPU server. The generated outputs are then communicated back to the robot, which enables it to describe the environment. We have also conducted some initial testing with the robot using real-life scenes. The results show that the robot can identify many objects and describe the environmental layouts and people accurately, however struggles with more complex scenes, owing to the longer processing time. In future work, we endeavor to enable the robot to react to scenes quicker and more accurately, which could be used for healthcare purposes, e.g. to alert care providers of a fallen person or if the fallen subject requires aid.

5 CONCLUSIONS

In this research, we have proposed a local based deep learning architecture for image description generation, in order to describe and annotate multiple objects and people within the given image. It consists of object detection and recognition, scene classification, RNN-based attribute prediction, and encoder-decoder RNNs for sentence generation. The proposed system mitigates the problems associated with holistic methods by relating specifically to image regions of people and objects in a given image in order to gain more detailed and longer descriptions. The experimental results have indicated the impressive performance of the proposed RNN-based object and human attribute prediction. Furthermore, the overall system also showed its significance for image description generation. Evaluated with the IAPR TC-12 dataset, in comparison with several baseline methods, i.e. NeuralTalk [8], NIC [9], Show, Attend and Tell [13] and Adaptive Attention [14], the proposed system produces more detailed and descriptive captions, and outperforms these state-of-the-art methods significantly for nearly all the evaluation metrics. The empirical results also indicated the superiority of our proposed system over existing methods when dealing with out-of-domain indoor scene description generation for images from the NYUv2 sentence dataset.

For future work, more detailed attributes, such as those related to garments as mentioned in [38], could be considered to further improve descriptive capabilities of the proposed system. We also aim to explore saliency detection to further improve the system's outputs with the emphasis of the potential focus of the images. In the longer term, we also aim to equip the proposed system with transfer learning [39] to deal with image description generation for images such as cartoons and oil paintings.

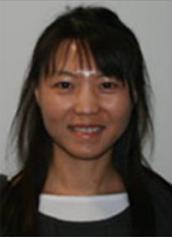
REFERENCES

- [1]. A. Oliva, A. Torralba, The role of context in object recognition. *Trends in cognitive sciences*, 11 (2007) 520-527.
- [2]. R. Girshick, J. Donahue, T. Darrell, J. Malik, Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proc. IEEE conf. Computer Vision and Pattern Recognition*. Columbus. Ohio. 2014. 580-587.
- [3]. P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R. Fergus, Y. LeCun, Overfeat: Integrated recognition, localization and detection using convolutional networks. In *Proc. Int. Conf. Learning Representations*. Banff. Canada. 2014.
- [4]. A. Farhadi, I. Endres, D. Hoiem, D. Forsyth, Describing objects by their attributes, In *Proc. IEEE conf. Computer Vision and Pattern Recognition*, Miami. Florida. 2009. 1778-1785
- [5]. S. Dhar, V. Ordonez, T.L. Berg, High level describable attributes for predicting aesthetics and interestingness. In *Proc. IEEE conf. Computer Vision and Pattern Recognition*. Colorado Springs., Colorado. 2011. 1657-1664
- [6]. L. Bourdev, S. Maji, J. Malik, Describing people: A poselet-based approach to attribute classification. In *Proc. IEEE Int. Conf. Computer Vision*, Barcelona., Spain. 2011. 1543-1550.
- [7]. H. Lang, H. Ling, Covert Photo Classification by Fusing Image Features and Visual Attributes. *IEEE Trans. Image Process.* 24 (2015) 2996-3008.
- [8]. A. Karpathy, F.F. Li, Deep visual-semantic alignments for generating image descriptions. In *Proc. IEEE conf. Computer Vision and Pattern Recognition*. Boston. Massachusetts. 2015. 3128-3137.
- [9]. O. Vinyals, A. Toshev, S. Bengio, D. Erhan, Show and tell: A neural image caption generator. In *Proc. IEEE conf. Computer Vision and Pattern Recognition*. Boston. Massachusetts. 2015. 3156-3164.
- [10]. D. Lin, C. Kong, S. Fidler, R. Urtasun, Generating multi-sentence lingual descriptions of indoor scenes. In *Proc. British Machine Vision Conference*. Swansea, UK. 2015.
- [11]. G. Kulkarni, V. Premraj, S. Dhar, S. Li, Y. Choi, A.C. Berg, T.L. Berg, Baby talk: Understanding and generating image descriptions. *IEEE Trans. Pattern Anal. Mach. Intell.* 35 (2011) 2891-2904
- [12]. A. Mathews, L. Xie, X. He, SentiCap: Generating Image Descriptions with Sentiments. In *Proc. Thirtieth AAAI Conference on Artificial Intelligence*. Phoenix. Arizona. 2015.
- [13]. K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhutdinov, R. Zemel, Y. Bengio, Show, Attend and Tell: Neural Image Caption Generation with Visual Attention. In *Proc. Conf. International Conference on Machine Learning*. 2015, 2048-2057.

- [14]. J. Lu, C. Xiong, D. Parikh, R. Socher, 2017. Knowing When to Look: Adaptive Attention via A Visual Sentinel for Image Captioning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.
- [15]. C. Liu, C. Wang, F. Sun and Y. Rui, 2016. Image2Text: A Multimodal Image Captioner. In Proceedings of the 2016 ACM on Multimedia Conference (pp. 746-748). ACM.
- [16]. J. Johnson, A. Karpathy and F.F. Li, 2016. Densecap: Fully convolutional localization networks for dense captioning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 4565-4574).
- [17]. S. Ren, K. He, R. Girshick and J. Sun, 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. In Advances in neural information processing systems (pp. 91-99).
- [18]. J. Mao, J. Huang, A. Toshev, O. Camburu, A.L. Yuille and K. Murphy, 2016. Generation and comprehension of unambiguous object descriptions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 11-20).
- [19]. Q. You, H. Jin, Z. Wang, C. Fang and J. Luo, 2016. Image captioning with semantic attention. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 4651-4659).
- [20]. H. Fang, S. Gupta, F. Iandola, R.K. Srivastava, From Captions to Visual Concepts and Back. In Proc. IEEE conf. Computer Vision and Pattern Recognition. Boston. Massachusetts. 2015, 1473-1482.
- [21]. Q. Wu, C. Shen, L. Liu, A. Dick, A.V.D. Hengal, What Value Do Explicit High Level Concepts Have in Vision to Language Problems. In Proc. IEEE conf, Computer Vision and Pattern Recognition. Las Vegas. Nevada, 2016. 203-212.
- [22]. O. Russakovsky, L. Fei-Fei, Attribute learning in large-scale datasets. In Trends and Topics in Computer Vision. 1-14. Springer Berlin Heidelberg. 2010.
- [23]. A. Krizhevsky, I. Sutskever, G.E. Hinton, Imagenet classification with deep convolutional neural networks. In Advances in neural information processing systems. 1097-1105. 2012.
- [24]. K.E.A. Van de Sande, J.R.R. Uijlings, T. Gevers, A.W.M. Smeulders, Segmentation as selective search for object recognition. In Proc. IEEE conf. Computer Vision and Pattern Recognition. Colorado Springs., Colorado. 2011. 1879-1886.
- [25]. D. Bahdanau, K. Cho, Y. Bengio, Neural machine translation by jointly learning to align and translate. arXiv preprint arXiv:1409.0473. Unpublished. 2014.
- [26]. A. Graves, J. Schmidhuber, Framewise phoneme classification with bidirectional LSTM and other neural network architectures. IEEE Trans. Neural Netw. 18 (2005) 602-610.
- [27]. N. Kumar, A.C. Berg, P.N. Belhumeur, S.K. Nayar, Attribute and simile classifiers for face verification. In Proc. IEEE Int. Conf. Computer Vision. Kyoto. Japan. 2009. 365-372.
- [28]. G.A. Miller, WordNet: a lexical database for English. Communications of the ACM, 38 (1995) 39-41.
- [29]. M. Grubinger, P. Clough, H. Müller, T. Deselaers, The IAPR TC-12 benchmark: A new evaluation resource for visual information systems. In Int. Workshop OntoImage. 13-23. 2006.
- [30]. T.Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, C.L. Zitnick, Microsoft coco: Common objects in context. In Computer Vision–ECCV 2014, 740-755.
- [31]. P. Young, A. Lai, M. Hodosh, J. Hockenmaier, From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions, Transactions of the Association for Computational Linguistics, 67-78. 2014.
- [32]. K. Papineni, S. Roukos, T. Ward, W.J. Zhu, BLEU: a method for automatic evaluation of machine translation. In Proc. of the 40th Annu. meeting on association for computational linguistics. ACL. 2002. 311-318.
- [33]. C.Y. Lin, Rouge: A package for automatic evaluation of summaries. In Proc. of the ACL-04 workshop on Text summarization branches out. 74-81. 2004.
- [34]. S. Banerjee, A. Lavie, METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In Proc. of the ACL workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization. 65-72. 2005.
- [35]. M. Denkowski, A. Lavie, Meteor universal: Language specific translation evaluation for any target language. In Proc. 9th Workshop on Statistical Machine Translation. 2014.
- [36]. C. Kong, D. Lin, M. Bansal, R. Urtasun, S. Fidler, What are you talking about? text-to-image coreference. In Proc. IEEE conf. Computer Vision and Pattern Recognition. 3558-3565. 2014.
- [37]. M. Everingham, L.V. Gool, C.K.I. Williams, J. Winn and A. Zisserman. The PASCAL Visual Object Classes (VOC) Challenge. International Journal of Computer Vision, 88(2), 303-338, 2010
- [38]. J. Shen, G. Liu, J. Chen, Y. Fang, J. Xie, Y. Yu, and S. Yan, Unified structured learning for simultaneous human pose estimation and garment attribute classification. IEEE Trans. Image Process. 23 (2014) 4786-4798.
- [39]. L. Shao, F. Zhu and X. Li, Transfer Learning for Visual Categorization: A Survey, IEEE Transactions on Neural Networks and Learning Systems, 26 (5) 1019-1034, 2015.



Philip Kinghorn received the BSc degree in Computer Science at Northumbria University, UK in 2014. He is currently a PhD student in the Department of Computer Science and Digital Technologies, Northumbria University. His research interests include computer vision, deep learning and intelligent robotics.



Li Zhang is an Associate Professor & Reader in Computer Science in Northumbria University, UK and also serving as an Honorary Research Fellow in the University of Birmingham, UK. Dr Zhang holds expertise in artificial intelligence, machine learning, intelligent robotics and affective computing. She also gained her PhD and postdoctoral experience from University of Birmingham previously. She has served as a programme co-chair and IPC member for international conferences and as an associate editor for Decision Support Systems. Dr Zhang is a member of IEEE.



Ling Shao is a professor with School of Computing Sciences at University of East Anglia. Previously, he was a professor (2014-2016) with Northumbria University, a senior lecturer (2009-2014) with University of Sheffield and a senior scientist (2005-2009) with Philips Research, The Netherlands. His research interests include computer vision, image/video processing and machine learning. He is an associate editor of IEEE Transactions on Image Processing, IEEE Transactions on Neural Networks and Learning Systems and several other journals. He is a Fellow of the British Computer Society and the Institution of Engineering and Technology. He is a senior member of IEEE.

ACCEPTED MANUSCRIPT