DECOUPLING CATEGORY-WISE INDEPENDENCE AND RELEVANCE WITH SELF-ATTENTION FOR MULTI-LABEL IMAGE CLASSIFICATION

Luchen Liu, Sheng Guo, Weilin Huang, Matthew R. Scott

Malong Technologies, Shenzhen, China
Shenzhen Malong Artificial Intelligence Research Center, Shenzhen, China

ABSTRACT

Multi-label image classification has achieved remarkable progress thanks to deep convolutional neural networks (CNNs). In this paper, we propose a Decouple Network (DecoupleNet) which is an end-to-end CNN-based framework able to trade off class-level feature independence and relevance during training. The proposed DecoupleNet is able to decouple category-wise independence and relevance with image-level supervision. We design a category-wise space-to-depth module with a spatial pooling strategy to exploit more meaningful convolutional features. They are integrated with class-wise correlated information which is automatically learned via a new self-attention mechanism. We conduct extensive experiments on two large-scale benchmarks: the MS-COCO and the NUS-WIDE, where the proposed DecoupleNet obtains impressive performance compared favorably against the state-of-the-art methods on multi-label image classification.

Index Terms— Multi-label image classification, self-attention, convolutional neural network

1. INTRODUCTION

Great successes have been achieved in image classification, due to the rapid development of deep convolutional neural networks (CNNs) [1, 2, 3]. Existing works mainly focus on single-label image classification problem by learning to assign a single class label to an image. However, for multi-label image classification task, an image often contains a set of labels with arbitrary number. Generally, multi-label classification can be transformed into a multiple binary classification problem. To explore the strong representation capability of CNNs, recent approaches aim to learn an end-to-end trainable model for multi-label classification. A straightforward approach is to fine-tune a typical image classification model for learning a set of class-wise binary classifiers. Besides, several recent CNN-based approaches aim to add extra structure in an effort to learn the relationship among various categories, by designing recurrent neural networks (RNNs) [4, 5, 6, 7], attention mechanism [8, 7], and ranking loss functions [9, 10]. With the development of weakly-supervised learning (WSL), a number of WSL-based methods [11, 12, 13] has obtained promising performance on multi-label image classification. These methods can achieve reasonable results based on the original CNN architectures, but the improvements were marginal by simply fine-tuning CNNs as multiple binary classifiers. These approaches do not explicitly exploit class correlation information, while deep networks are capable of modeling such relationship implicitly. For example, CNN-based binary classifiers can learn class relevance by using convolutional filters which are applied to different channels. Adding extra components to the original CNN architectures has potential to improve the performance by learning stronger relationship between classes, but it would result in an increased learning difficulty in training, and could inevitably cause information redundancy to some extent.

There has been a number of recent works focusing on exploring correlations of deep features within CNN architectures [14, 15, 16, 8, 17, 18, 19]. For example, Network-in-Network [20] uses standard convolutional filters to jointly learn feature combinations. In [16], cross-channel correlations are modeled by independent spatial structures. Furthermore, attention mechanism is widely-used to explore the correlations of CNN features. In [8, 17], spatial correlations can be modeled by learning a designed attention. Wang *et al*. [18] proposed Residual Attention Networks by designing a stacking attention model to learn attention-aware features, and Hu *et al.* proposed Squeeze-and-Excitation Networks (SENet) able to learn channel-wise attention maps that encode cross-channel correlations [19]. Inspired by these approaches, our goal is to design an automatic learning mechanism able to explicitly exploit the strong relationship between classes, rather than the implicit category-wise relevance learned by the original network itself. Both class-independent features and class-correlated features are important to learning multi-label classifiers. The class-correlated information can be learned by using attentional maps, and then improves multi-label classification performance when incorporated properly into the class-independent features.

To this end, we propose a Decouple Network (DecoupleNet) which is an end-to-end trainable CNN-based approach that directly learns class independence and category-

Weilin Huang is the corresponding author (whuang@malong.com).

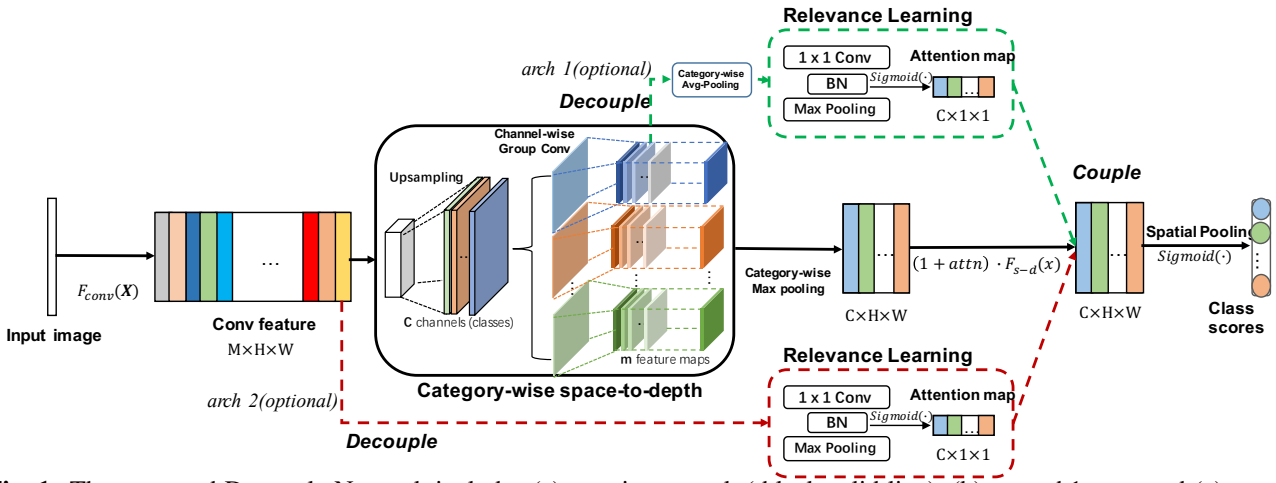

Fig. 1. The proposed Decouple Network includes (a) a main network (black solid line), (b) an arch1 part, and (c) an arch 2 part. The arch1 and arch2 can learn class-wise relevance attention maps from the independent multiple features of categories (green dotted line) and the backbone features (red dotted line), respectively.

wise relevance in the convolutional features. We introduce a new self-attention mechanism that allows class independence and relevance to work collaboratively in the multi-label training process. We design a category-wise space-to-depth module which is integrated seamlessly into the original CNN framework. This allows us to preserve more detailed features, and enhance class independence in the main network. In addition, upsampling and channel-wise convolution are applied jointly to learn multiple feature maps for each class, which enhances class independence. Finally, we conduct extensive experiments and evaluations on two multi-label benchmarks: the MS-COCO and the NUS-WIDE. Experimental results show that our methods outperform existing state-of-the-art methods. Visualization and analysis further demonstrate the effectiveness of our methods comprehensively.

2. DECOUPLE NETWORK

Overview. The proposed Decomple Network is shown in Fig. 1. Our framework is composed of two sub-networks: a main network and an attention sub-network. The main network is able to learn class-independent features with image-level supervision, and the attention sub-network has an attention layer capable of learning the correlated relationship between classes. We use a pre-trained backbone convolutional network for feature extraction (which is also involved in training), and then the convolutional features are encoded into class-wise features for classification via the main network. Meanwhile, the attention sub-network learns channel-wise attention maps, which are then integrated into the class-wise features.

2.1. Main-Net for Enhancing Class-Wise Independence

We design a main network capable of enhancing class-independent features in the convolutional layers. It consists of an upsampling layer, a class-independent multi-map learning layer and a category-wise max-pooling layer.

Upsampling Operation We introduce an upsampling operation to preserve more local detailed information in convolutional maps, by increasing spatial resolution. The feature maps are up-sampled via a 3×3 transposed convolutional kernel. The input features are encoded into C channels, where C is the number of classes via the upsampling layer. Each channel preserves spatial details for each category.

Space-to-Depth Structure To further enhance class-wise independence, we design a class-wise space-to-depth encoder structure. We apply a 3×3 channel-wise group convolution that encodes each upsampled channel independently, and each filter can only learns features from a specific channel. At the same time, the spatial size of each feature is reduced. By this operation, the spatial information is transformed into m multiple maps, and each map can represent different spatial features. We apply a category-wise max-pooling for combining the discriminative information among m class-wise feature maps, and decode them into a single map with a same spatial resolution. It transforms CNN features into C class-wise convolutional features, allowing each category to focus on its own channels, which enhance independence and discriminative power of the learned deep representation.

2.2. Attention Sub-Net for Relevance Learning

The attention sub-network is designed to decouple class-wise independence and relevance. As shown in Fig. 1, we design two optional decoupled operations. The first one is to learn class relevance directly from the class-independent convolutional features - category-wise multiple convolutional maps. This allows the proposed self-attention layer to learn the class relevance information from class-independent features, which in turn compensate each other. The second one can learn class

relevance from the backbone convolutional features, which may be able to preserve more local detailed information in the main network.

Self-Attention Learning It can be considered as a dynamic re-weighting operation that automatically learns channel-wise attention maps without additional supervision. Inspired by SENet [19], we design an attention layer to model the interdependencies between different channels of the convolutional features. Specifically, the attention layer is composed of a 1×1 convolution layer, a batch normalization layer and a max-pooling layer. We use a global max-pooling with *Sigmoid* to learn channel-wise attention maps. The proposed self-attention layer adaptively models the class-wise relevance, without any additional explicit supervision. To integrate such class relevance and the learned class-wise independence, we use a residual connection to encode the learned attention maps into the convolutional features of the main network. For input maps X_{in}, we can have

$$attn = U(X_{in}; W), \;\; attn \in \mathbb{R}^{C\times1\times1}, \tag{1}$$

$$X_{coupled} = (1 + attn) \cdot Z, \;\; X_{coupled} \in \mathbb{R}^{C\times14\times14} \tag{2}$$

where $U(\cdot)$ is the attention layer. $X_{coupled}$ is the feature maps coupled with the learned attention, and Z is the features learned by the main-net.

Spatial Pooling To integrate the capabilities of max-pooling and global average pooling, a spatial pooling method was proposed in [12], where optional hyper-parameters are used to average both positive and negative pixels in each feature map,

$$s^c = \frac{1}{k^+} \sum_{topk^+ p^c_{i,j}} p^c_{i,j} + \alpha \frac{1}{k^-} \sum_{topk^- p^c_{i,j}} p^c_{i,j} \tag{3}$$

where s^c is the score of channel c, and $p^c_{i,j}$ is a pixel of the c-th feature map $X_{coupled}$. k^+ is the number of pixels with the highest values, and k^- for the lowest values. By using a *Sigmoid* layer, we obtain a final probability score for each class. We apply a Binary Cross-entropy loss in the training.

3. EXPERIMENTAL RESULTS AND COMPARISONS

3.1. Evaluation Metrics

By following the standard evaluation metrics on multi-label classification, we use mean average precision (mAP) in VOC2012[1] to evaluate all methods. We employ a macro/micro precision (P_C/P_O), a macro/micro recall (R_C/R_O) and a macro/micro F1-measure ($F1_C$/$F1_O$). The "macro" means that it is evaluated by averaging per-class metric values, and"micro" indicates that it is an overall measure for all images. We empirically report the labels having a score over 0.5. Each input image is resized to 448×448. We set two strong baselines by using ResNet101 [2] and WILDCAT [12],

[1]http://host.robots.ox.ac.uk/pascal/VOC/voc2012

Table 1. Results on the MS-COCO Dataset.

Method	mAP	P-C	R-C	F1-C	P-O	R-O	F1-O
WARP	-	59.3	52.5	55.7	59.8	61.4	60.7
CNN-RNN	-	66.0	55.6	60.4	69.2	66.4	67.8
RLSD	68.2	67.6	57.2	62.0	70.1	63.4	66.5
RNN-RL	-	78.8	57.2	66.2	84.0	61.6	71.1
OF-RNN (w/ attn)	-	71.6	54.8	62.1	74.2	62.2	67.7
RNN-Attention	73.4	79.1	58.7	67.4	84.0	63.0	72.0
ResNet101-SRN	77.1	81.6	65.4	71.2	82.7	69.9	75.8
ResNet101-MEFF	-	80.4	70.2	74.9	85.2	72.5	78.4
ResNet101(baseline)	77.8	80.7	65.9	71.8	84.3	71.0	77.1
WILDCAT	80.7	81.3	70.0	74.8	**85.3**	73.5	78.9
DecoupleNet(arch1)	81.7	82.9	70.8	75.6	85.0	73.5	78.8
DecoupleNet(arch2)	81.8	82.3	71.1	75.7	84.3	74.2	78.9
DecoupleNet(arch1+arch2)	**82.2**	**83.1**	**71.6**	**76.3**	84.7	**74.8**	**79.5**

Table 2. Results on the NUS-WIDE Dataset.

Method	mAP	P-C	R-C	F1-C	P-O	R-O	F1-O
KNN	-	32.6	19.3	24.3	43.9	53.4	47.6
Softmax	-	31.7	31.2	31.4	47.8	59.5	53.0
WARP	-	31.7	35.6	33.5	48.6	60.5	53.9
CNN-RNN	-	40.5	30.4	34.7	49.9	61.7	55.2
RLSD	54.1	44.4	49.6	46.9	54.4	67.6	60.3
OF-RNN(w/ attn)	-	**59.4**	50.7	54.7	69.0	**71.4**	70.2
ResNet101(baseline)	52.4	55.4	48.9	49.4	69.4	70.3	69.8
WILDCAT	52.3	52.4	51.5	50.6	70.0	69.2	69.6
DecoupleNet(arch1)	56.8	56.5	**55.9**	53.9	69.8	70.4	70.1
DecoupleNet(arch2)	56.9	59.0	54.0	53.9	70.8	70.0	70.4
DecoupleNet(arch1+arch2)	**57.5**	58.5	55.3	**54.8**	**70.9**	70.6	**70.7**

where WILDCAT is a weakly-supervised method for multi-label classification and localization.

3.2. Results

MS-COCO [21] is a widely-used dataset with 80 objective labels. The training set contains 82,787 images, and all experiments are tested on the validation set which has 40,504 images. In addition to two baselines by using ResNet101 and WILDCAT, we further compare DecoupleNet with recent state-of-the-art approaches, including WARP [9], CNN-RNN [4], RLSD [22], RNN-Attention [7], RNN-RL [23], OF-RNN [6], SRN [8] and MEFF [24]. Note that SRN and MEFF use the same ResNet101 backbone as our models, while ResNet152 is used by OF-RNN as backbone. The comparisons are summarized in Table 1. As can be found, DecoupleNet outperforms the state-of-the-art methods in most metrics on the MS-COCO. It has an improvement of 1.5% mAP over the best baseline, WILDCAT. Specifically, both arch2 and arch1 can improve the performance of ResNet101 baseline by a large margin - 4% mAP. The performance of our ensemble model is significantly higher than

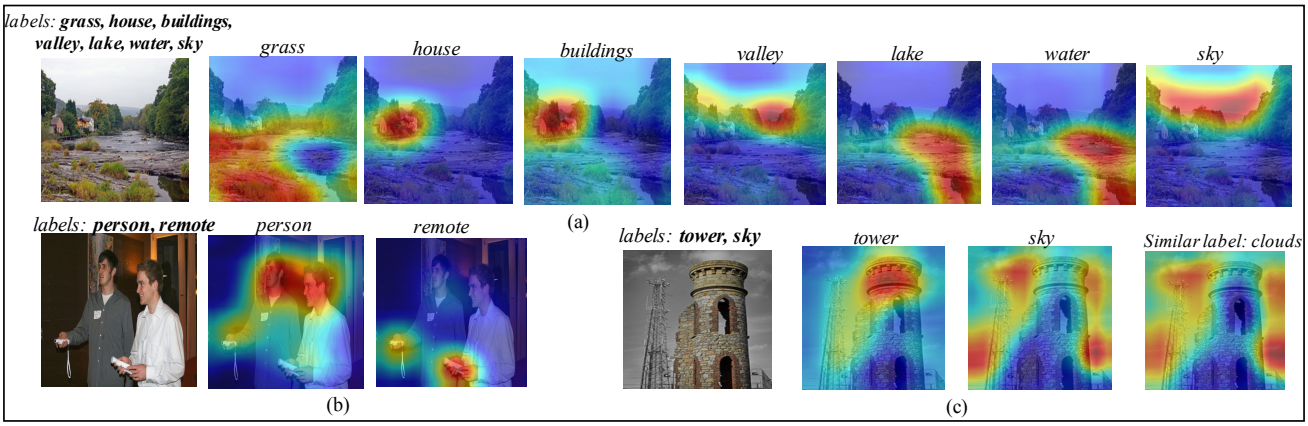

Fig. 2. Visualization of heat-maps for corresponding category-wise feature maps in image examples. The spatial localization and tiny object are shown in (a) and (b). Conceptions as *house* and *buildings*, *lake* and *water* are shown in (a), even similar semantic labels which are not included in the ground truth are learned as shown in (c).

that of ResNet101-based methods compared, leading to an improvement of about 4% in the term of mAP.

NUS-WIDE [25] contains 269,648 images with 81 concepts collected from Flickr. We use the official train/test split, where 161,789 images are used for training, and 107,859 images are used for test. To further evaluate our methods, we compares DecoupleNet with several state-of-the-art approaches, such as WARP [9], RLSD [22], CNN-RNN [4] and OF-RNN [6]. As shown in Table 2, DecoupleNet can obtain consistent performance improvements over the compared approaches. Specifically, both the proposed architectures outperform two baseline methods significantly with a large margin of ∼4.5% mAP. Even the OF-RNN uses a deeper backbone - ResNet152 , DecoupleNets are still compared favorably against it. An ensemble of arch1 and arch2 has clear improvements in the term of mAP, F1-C and F1-O.

Table 3. Comparisons with/without class-wise relevance on the NUS-WIDE via different backbones.

Method	mAP	
	ResNet50	ResNet101
DecoupleNet (w/o attn)	56.0	56.4
DecoupleNet (arch1)	**56.5**	56.8
DecoupleNet (arch2)	**56.5**	**56.9**

Visualization We visualize convolutional feature maps of the category-wise max-pooling layer to demonstrate the discriminative regions learned by the proposed methods. As shown in Fig. 2, the proposed method is able to learn rough attention regions of the key category information in the corresponding class-wise feature maps. It is also able to learn similar semantic conceptions and class-wise correlations.

Table 4. Comparisons of different upsampling strategies on the MS-COCO. *arch1* and ResNet50 backbone are used.

Backbone	Method	mAP	F1-C	F1-O
ResNet50	Ours (w/o upsampling)	77.7	70.7	75.7
	Ours (bilinear)	79.5	73.0	**77.5**
	Ours (transposed conv)	**80.0**	**73.7**	77.4

3.3. Ablation Study

Impact of Attention Sub-Network We train our models by removing the attention sub-network branch (without the class-wise attention layer). In Table 3, it can be observed that improvement gains can be achieved on the NUS-WIDE when the attention sub-network is applied, with an improvement of about 0.5% mAP for two backbones.

Impact of Upsampling We further investigate different upsampling methods, such as transposed convolution, bilinear upsampling, and evaluate the efficiency of the upsampling layer in Table 4. The results demonstrate that the upsampling operation is efficient, and the transposed convolution performs better than the bilinear one.

4. CONCLUSION

We have presented new decouple networks for multi-label image classification, where a novel category-wise space-to-depth module with spatial pooling strategy is proposed to exploit more meaningful class-independent convolutional features. Then we design an attention sub-network able to learn the correlated relationship between classes via a self-attention mechanism. Both components are integrated seamlessly into a single end-to-end trainable model. The proposed DecoupleNet outperforms the state-of-the-art methods on the MS-COCO and NUS-WIDE datasets, and visualization and analysis further confirm its effectiveness.

5. REFERENCES

[1] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton, "Imagenet classification with deep convolutional neural networks," in *NIPS*, 2012, pp. 1097–1105.

[2] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *CVPR*, June 2016.

[3] Limin Wang, Sheng Guo, Weilin Huang, Yuanjun Xiong, and Yu Qiao, "Knowledge guided disambiguation for large-scale scene classification with multi-resolution cnns," *IEEE Transactions on Image Processing*, vol. 26, no. 4, pp. 2055–2068, 2017.

[4] Jiang Wang, Yi Yang, Junhua Mao, Zhiheng Huang, Chang Huang, and Wei Xu, "Cnn-rnn: A unified framework for multi-label image classification," in *CVPR*, June 2016.

[5] Yao-Yuan Yang, Yi-An Lin, Hong-Min Chu, and Hsuan-Tien Lin, "Deep learning with a rethinking structure for multi-label classification," *CoRR*, vol. abs/1802.01697, 2018.

[6] Shang-Fu Chen, Yi-Chen Chen, Chih-Kuan Yeh, and Yu-Chiang Frank Wang, "Order-free RNN with visual attention for multi-label classification," *CoRR*, vol. abs/1707.05495, 2017.

[7] Zhouxia Wang, Tianshui Chen, Guanbin Li, Ruijia Xu, and Liang Lin, "Multi-label image recognition by recurrently discovering attentional regions," in *ICCV*, Oct 2017.

[8] Feng Zhu, Hongsheng Li, Wanli Ouyang, Nenghai Yu, and Xiaogang Wang, "Learning spatial regularization with image-level supervisions for multi-label image classification," in *CVPR*, July 2017.

[9] Yunchao Gong, Yangqing Jia, Thomas Leung, Alexander Toshev, and Sergey Ioffe, "Deep convolutional ranking for multilabel image annotation," *arXiv preprint arXiv:1312.4894*, 2013.

[10] Yuncheng Li, Yale Song, and Jiebo Luo, "Improving pairwise ranking for multi-label image classification," in *CVPR*, July 2017.

[11] Thibaut Durand, Nicolas Thome, and Matthieu Cord, "Weldon: Weakly supervised learning of deep convolutional neural networks," in *CVPR*, June 2016.

[12] Thibaut Durand, Taylor Mordan, Nicolas Thome, and Matthieu Cord, "Wildcat: Weakly supervised learning of deep convnets for image classification, pointwise localization and segmentation," in *CVPR*, July 2017.

[13] Sheng Guo, Weilin Huang, Haozhi Zhang, Chenfan Zhuang, Dengke Dong, Matthew R Scott, and Dinglong Huang, "Curriculumnet: Weakly supervised learning from large-scale web images," in *ECCV*, 2018, pp. 135–150.

[14] Sheng Guo, Weilin Huang, Limin Wang, and Yu Qiao, "Locally supervised deep hybrid model for scene recognition," *IEEE transactions on image processing*, vol. 26, no. 2, pp. 808–820, 2017.

[15] Jimmy Ba, Volodymyr Mnih, and Koray Kavukcuoglu, "Multiple object recognition with visual attention," *CoRR*, vol. abs/1412.7755, 2014.

[16] Francois Chollet, "Xception: Deep learning with depthwise separable convolutions," in *CVPR*, July 2017.

[17] Han Hu, Jiayuan Gu, Zheng Zhang, Jifeng Dai, and Yichen Wei, "Relation networks for object detection," in *CVPR*, June 2018.

[18] Fei Wang, Mengqing Jiang, Chen Qian, Shuo Yang, Cheng Li, Honggang Zhang, Xiaogang Wang, and Xiaoou Tang, "Residual attention network for image classification," in *CVPR*, July 2017.

[19] Jie Hu, Li Shen, and Gang Sun, "Squeeze-and-excitation networks," in *CVPR*, 2018.

[20] Min Lin, Qiang Chen, and Shuicheng Yan, "Network in network," *arXiv preprint arXiv:1312.4400*, 2013.

[21] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick, "Microsoft COCO: common objects in context," in *ECCV*, 2014, pp. 740–755.

[22] Junjie Zhang, Qi Wu, Chunhua Shen, Jian Zhang, and Jianfeng Lu, "Multi-label image classification with regional latent semantic dependencies," *CoRR*, vol. abs/1612.01082, 2016.

[23] Tianshui Chen, Zhouxia Wang, Guanbin Li, and Liang Lin, "Recurrent attentional reinforcement learning for multi-label image recognition," in *AAAI*, 2018.

[24] Weifeng Ge, Sibei Yang, and Yizhou Yu, "Multi-evidence filtering and fusion for multi-label classification, object detection and semantic segmentation based on weakly supervised learning," in *CVPR*, June 2018.

[25] Tat-Seng Chua, Jinhui Tang, Richang Hong, Haojie Li, Zhiping Luo, and Yantao Zheng, "NUS-WIDE: a real-world web image database from national university of singapore," in *CIVR*, 2009.