View and illumination invariant object classification based on 3D Color Histogram using Convolutional Neural Networks

Earnest Paul Ijjina and C. Krishna Mohan

Indian Institute of Technology Hyderabad
Yeddumailaram, Telangana, India 502205
cs12p1002,ckm@iith.ac.in

Abstract. Object classification is an important step in visual recognition and semantic analysis of visual content. In this paper, we propose a method for classification of objects that is invariant to illumination color, illumination direction and viewpoint based on 3D color histogram. A 3D color histogram of an image is represented as a 2D image, to capture the color composition while preserving the neighborhood information of color bins, to realize the necessary visual cues for classification of objects. Also, the ability of convolutional neural network (CNN) to learn invariant visual patterns is exploited for object classification. The efficacy of the proposed method is demonstrated on Amsterdam Library of Object Images (ALOI) dataset captured under various illumination conditions and angles-of-view.

1 Introduction

Object recognition is an active area of research for the last five decades [1] and efficient recognition of objects under varying illumination conditions is a problem yet to be solved. Visual object classification is an important step in visual recognition and semantic analysis of visual content. Some of the challenges in object classification from 2D images is the loss of depth information, variations in the visual information captured, due to the change in view-angle, illumination color and illumination direction of the object. Techniques relying solely on shape and texture features are computationally expensive and inefficient to classify non-symmetric objects with complex shape. Thus, features incorporating color information should be considered for the design of an efficient object classification mechanism robust to variations in object viewpoint and illumination conditions.

Representing object images as graphs built over corner point, is proposed in [2] to classify objects using graph matching. The effectiveness of this approach depends upon the robustness of the graph representation against varying illumination conditions and angles of view. The dependence of this approach on corner points makes the representation less distinctive and thereby affects the efficiency. To compute similarity of images at multiple resolutions effectively, a new multi-resolution distance metric, Manhattan-pyramid distance is proposed in [3]. A multiset discriminant canonical correlation method namely multiple principle angle [4] that iteratively learns multiple subspaces and the global discriminant subspace to consider both local and global canonical correlations is used to classify images of objects captured from different view angles. For object classification, GMMs based on views of each object are built in [5] from global models using maximum likelihood estimation followed by an adaptation step to minimize the kNN classification error rate. The GMMs are combined to minimize the distance between objects of the same class and maximize the distance between objects of different classes. This method may not be effective for illumination invariant object recognition due to its dependence on shape information. An attention guided model for object recognition is proposed in [6] that learns the probability of an objects visual appearance having a range of values within a particular feature map. For a given test image, the possible candidate classes were identified along with their probabilities. As color is one of the critical feature, the model encounters more ambiguity when classifying images with variation in illumination conditions.

Label consistent K-SVD algorithm [7] associates label information with each dictionary item to learn a discriminative dictionary for object classification using spatial pyramid features. Multiple sets of features are combined using multiple kernel learning (MKL) for object recognition [8]. Representation learning algorithms [9] like deep learning, autoencoder and deep networks that learn from generic priors are used for object representations and classification. Convolutional neural network based feature extraction and classification models [10] [11] trained on ImageNet dataset, produced competitive results for localization, detection and classification tasks for the last two years.

Some of the limitations of existing approaches is their lack of robustness to illumination variation and high computationally complexity. In this paper, we propose a view and illumination invariant object classification system using a convolutional neural network. The reminder of this paper is organized as follows: In Section 2, the proposed approach, image representation and the object classifier are discussed. Experimental results were discussed in Section 3. Section 4 gives conclusion of this work.

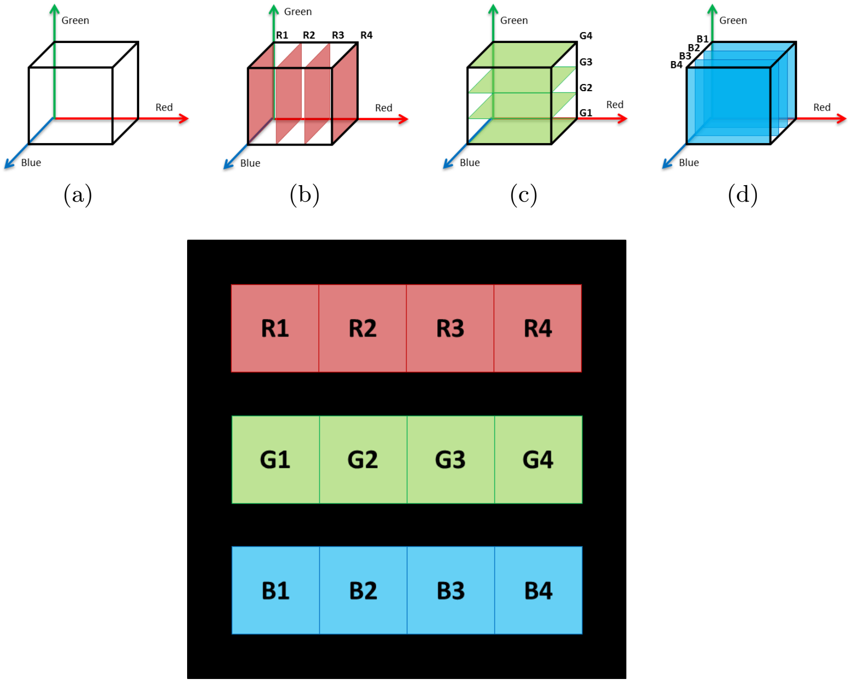
2 Proposed Approach

In this work, we propose a method for view and illumination invariant object classification based on 3D color histogram information using a convolutional neural network (CNN). The steps involved in the generation of 2D representation are detailed in the following section.

2.1 Image representation

Due to the robustness of color distribution against changes in viewpoint and illumination conditions, a 2D representation preserving the neighborhood information of color bins in 3D color histogram of an image is used as the feature.

This study considers RGB 3D histogram with n bins (considering 4 for illustration) along each axis, resulting in a 4×4×4 cube with each axis representing a color in RGB as shown in Fig.1 (a). Fig.1 (b) shows the slices of the cube R1, R2, R3, R4 along red-axis. Similarly, the slices along green axis G1, G2, G3, G4 and along blue axis B1, B2, B3, B4 are shown in Fig.1 (c) and 1 (d). These 4×4 red, green and blue slices are arranged in a 20×20 matrix with a margin of 2 elements from each border and between slices of different color, to construct the 2D representation of an image, as shown in Fig.1 (e).



(a) (b) (c) (d)

(e)

Fig. 1. 2D representation of an RGB 3D color histogram. (a) cube representing 3D color histogram (b) slices of cube along red-axis (c) slices of cube along green-axis (d) slices of cube along blue-axis (e) proposed 2D representation of a 3D color histogram

The proposed 2D representation preserves the neighborhood information of color bins along an axis in slices along the remaining axis. The 2D representation of some objects in ALOI dataset is shown in Fig.2.

From the 2D representation of objects in Fig.2, it can be observed that the local patterns provide discriminative information useful for classification. Different images of the same object may have different patterns due to variations in image capturing conditions like view-angle, illumination color and illumination direction of the object. The classifier employed to capture the variations in local

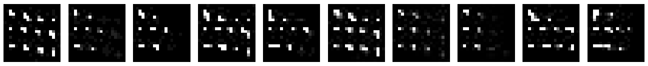


Fig. 2. The 2D representation of images of some ALOI objects

patterns of each object for effective classification, is elaborated in the following section.

2.2 Object classification using CNN

A convolutional neural network (CNN) [12] is a feed-forward neural network capable of recognizing local patterns with some degree of shift and distortion. This characteristic is explored to classify objects from the local patterns in their color histogram. The convolutional neural networks (CNNs) have been shown to outperform the standard fully connected deep neural networks in various computer vision challenges. A typical CNN architecture used as a classifier [13] consists of an alternating sequence of convolution and subsampling layers followed by a neural network for classification. The architecture used in the proposed approach is shown in Fig.3. If we consider a 20×20 image as input, 2×2 mask in subsampling layers S1 & S2, a 5×5 mask in convolution layers C1 & C2, 15 feature maps in F1 & F2 and 30 feature maps in F3 & F4, then 1) F1 represents 15 feature maps of size 16×16 2) F2 represents 15 feature maps of size 8×8 3) F3 represents 30 feature maps of size 4×4 and 4) F4 represents 30 feature maps of size 2×2. The 120 feature (corresponding to F4) generated at the output of second subsampling layer are given as input (I) to a fully connected neural network, to generate an output O to classify the N objects.

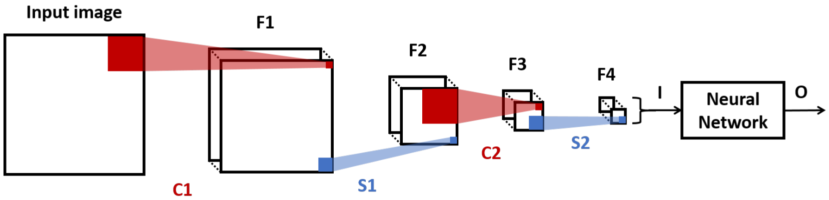


Fig. 3. CNN architecture in the proposed approach

The CNN is trained using back-propagation algorithm in batch-mode for 2500 epochs on training dataset and tested on test dataset to classify objects from the 2D representation of their images. Two-fold cross validation is used to evaluate the performance of the proposed approach. The train and test datasets are obtained by arranging the images of an object in order and assigning the odd image to the train dataset and even to the test dataset.

3 Experimental results

The proposed approach is evaluated on vegetable and fruit objects in ALIO dataset [14], that consists of images of 1000 objects captured under 24 configurations of illumination direction, 12 configuration of illumination color and 72 directions of object view point. As two-fold cross validation is used for evaluation, the initial train and test datasets are obtained by considering all the odd images of an object as training images and even as testing images. The train and test datasets are interchanged during cross-validation. The vegetable and fruit objects in ALOI dataset considered for evaluation are listed in Tables 1 and 2.

Table 1. Vegetables

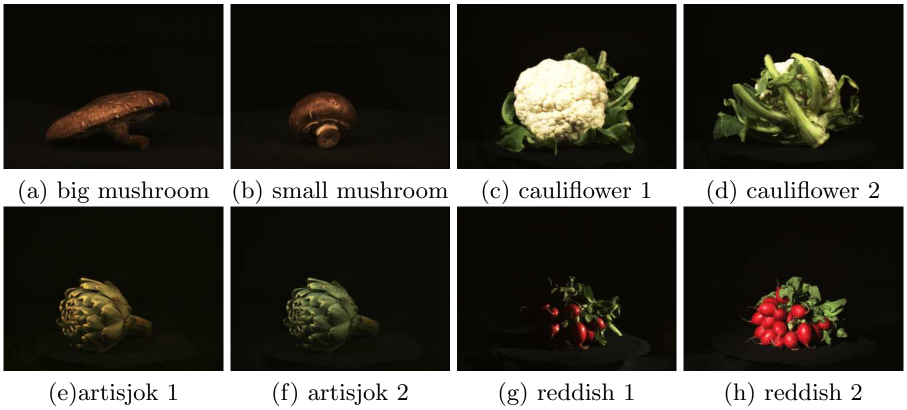
#	ALOI #	Object Name
1	17	Red onion
2	281	garlic
3	287	red onion
4	324	red pepper
5	709	big mushroom
6	711	small mushroom
7	714	onion
8	717	small onion
9	718	garlic
10	719	tomato
11	720	red onion
12	723	flat french bean
13	724	french bean
14	877	cauliflower
15	880	carrot
16	881	courgette
17	883	asperges
18	884	rettig
19	885	sweet potato
20	887	witlof
21	889	egg plant
22	948	green capsicum
23	952	artisjok
24	953	yellow capsicum
25	954	reddish

Table 2. Fruits

#	ALOI #	Object Name
1	3	Apricot
2	52	hairy ball
3	69	tomato
4	82	Apple
5	102	Kiwi
6	273	lemon2
7	446	orange
8	567	Pear
9	649	apple
10	650	kiwi
11	651	lemon
12	705	green capsicum
13	706	red capsicum
14	707	mango
15	708	kiwi
16	710	apple
17	712	mandarin
18	713	lemon
19	715	Unknown fruit
20	716	Unknown fruit 2
21	721	pear
22	722	lemon
23	870	pineapple
24	873	mango
25	879	cucumber
26	882	orange
27	888	melon
28	947	sherry tomatos
29	950	banana's

3.1 ALOI vegetable objects

Some examples of vegetable object images considered in this evaluation are shown in Fig.4. The Figs 4(a), 4(b) show the inter-class similarity of objects; Figs 4(c), 4(d) the intra-class dissimilarity of an object in terms of color profile. The Figs 4(e), 4(f) are images of the same object captured under different direction of illumination and Figs 4(g), 4(h) are images of the same object with different illumination temperate. This inter-class similarity and intra-class diversity makes vegetable categorization a challenging task.


(a) big mushroom (b) small mushroom (c) cauliflower 1 (d) cauliflower 2

(e)artisjok 1 (f) artisjok 2 (g) reddish 1 (h) reddish 2

Fig. 4. Examples images of ALOI vegetable objects

The 25 ALOI vegetable objects listed in Table 1 are used for evaluation. The CNN classifier is trained in batch mode with a batch size of 90 and evaluated on the test dataset. The following section presents the impact of the number of color bins used in 2D representation, the number of feature maps considered and the size of mask used in convolution layers.

Impact of configuration changes We assume that same size of convolution mask is used in both convolution layers and that the number of feature maps in F3 & F4 is double the number of feature maps in F1 & F2. The impact of number of bins used in the generation of color histogram, the size of the mask used in convolution layers and the number of feature maps used in the first convolution layer on the performance of vegetable object classification is presented in Table 3.

From Table 3, it can be observed that when the number of color bins is 3 or 4, increase in the number of feature maps generally improves the performance. When the number of color bins is 5 or 6, increase in number of feature maps deteriorates the performance. This suggests that a solution with optimal performance and time-complexity can be identified by considering the right set of

Table 3. Average classification error of CNN classifier (in %) for vegetable objects with execution time/iteration in parenthesis

# of bins	3 × 3 mask			5 × 5 mask			7 × 7 mask		
	5FM	15FM	25FM	5FM	15FM	25FM	5FM	15FM	25FM
3	12.6 (3.4s)	6.9 (22.2s)	5.7 (58s)	11.3 (4s)	4.3 (25s)	4.8 (64s)			
4	4.2 (4.4s)	3.4 (29s)	1.3 (78s)	1.8 (4.9s)	1.2 (31.4s)	1.4 (86s)	2.2 (5.9s)	1.4 (34s)	1.3 (87s)
5	1.1 (7s)	0.96 (45.8s)	76 (120s)	1.3 (9.5s)	0.74 (60s)	60 (152s)	0.59 (12.6s)	0.74 (77s)	1.0 (195s)
6	0.74 (8.9s)	80 (55s)	83 (142s)	0.81 (13.6s)	48 (89.5s)	92 (230s)	0.74 (16.2s)	1.0 (102.5s)	92 (276.5s)

values for these parameters. The computation time per iteration (5 epochs) also increases with the increase in number of feature maps, due to the increase in number of free variables to be tuned by the back-propagation algorithm. The confusion matrix of the trained classifier for images of vegetable category is shown in Fig.5.

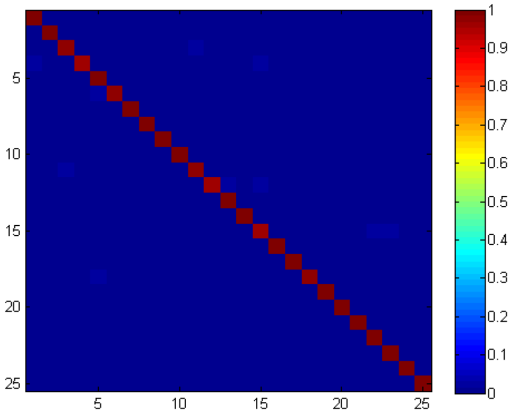
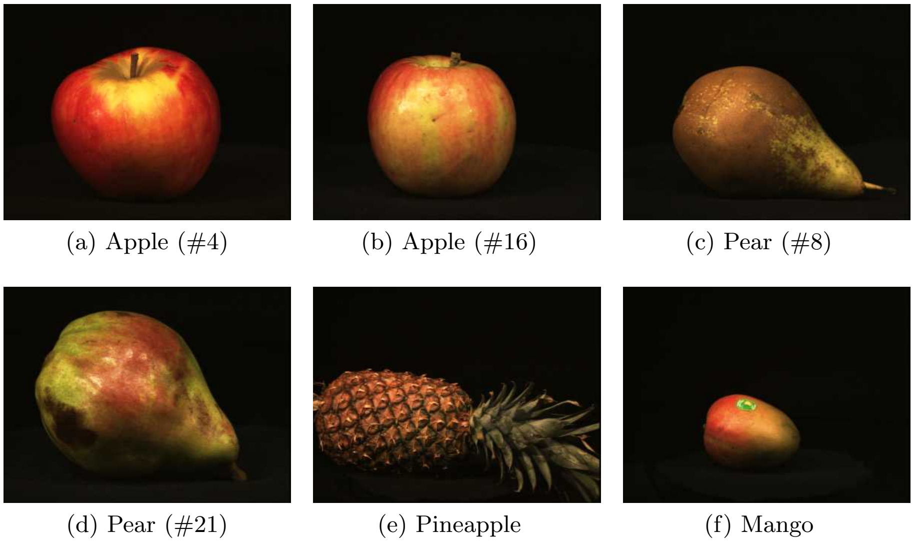


Fig. 5. Confusion matrix of vegetable object images from ALOI dataset

The class labels from top-left to top-right and bottom-left are in the order given in Table 1. The average classification error of our approach using 2-fold cross validation for vegetable objects is 0.74%.

3.2 ALOI fruit objects

Among the 29 ALOI fruit objects considered, there are multiple instances of the same object class like for apple, pear, kiwi etc., capturing the raw and ripe variants of these fruits, thereby making this a fine-grained classification. Some example fruit images are shown in Fig. 6, where Figs 6(a), 6(b) and 6(c), 6(d) are multiple instances of the same object i.e., apple and pear respectively, with different color profiles. The images in Figs. 6(e) and 6(f) show the complex texture and color of fruits being recognized. Thus, fruit classification is relatively more challenging than vegetable classification due to their complex texture and variants.


(a) Apple (#4) (b) Apple (#16) (c) Pear (#8)

(d) Pear (#21) (e) Pineapple (f) Mango

Fig. 6. Examples images of ALIO fruit objects

The CNN classifier is trained on the 29 fruit objects with a batch size of 87. The following section presents the impact of the number of color bins used in 2D representation, the number of feature maps considered and the size of mask used in convolution layers.

Impact of configuration changes We consider the same assumptions on the configuration of CNN architecture as we did for the classification of vegetable objects. The affect of number of bins used in the generation of color histogram, the size of the mask used in convolution layers and the number of feature maps used in the first convolution layer on the performance of fruit object classification is shown in Table 4.

Table 4. Average classification error of CNN classifier (in %) for fruit objects

# of bins	3 × 3 mask 5FM	 15FM	 25FM	5 × 5 mask 5FM	 15FM	 25FM	7 × 7 mask 5FM	 15FM	 25FM
3	7.21	4.85	4.34	6.25	3.57	3.7			
4	3.95	3	3	3.83	2.29	2.49	2.74	2.49	1.85
5	1.59	52.87	93.1	1.91	1.66	66.21	2.49	1.59	42.4
6	1.53	69.15	96.55	1.66	86.2	89.14	1.66	1.47	48

The confusion matrix of the trained classifier for images of vegetable objects is shown in Fig. 7.

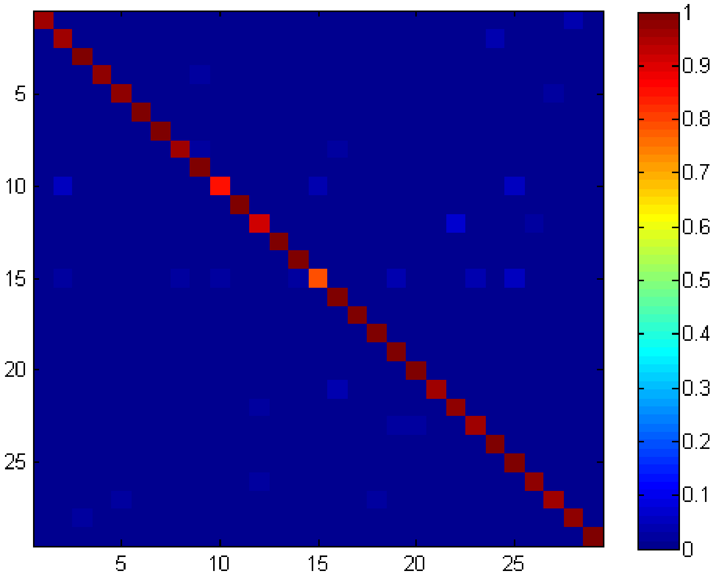
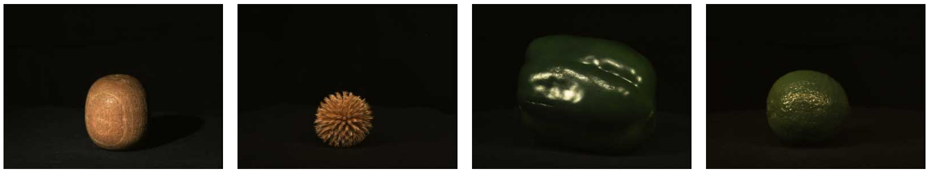


Fig. 7. Confusion matrix of fruit objects images from ALOI dataset

The class labels are as listed in Table 2. The average classification error of the proposed approach using 2-fold cross validation is 2.17%.

3.3 Analysis of results and comments

The low misclassification error shown in Figs 5 and 7 is due to existence of similar color profile for some objects under certain image capturing conditions. Fig. 8 shows some examples of objects Figs 8(a), 8(c) misclassified as objects of Figs 8(b), 8(d).



(a) Kiwi (10) (b)hairy ball (2) (c)capsicum(12) (d)lemon (22)

Fig. 8. Some misclassified fruit objects

The classification accuracy of ALOI vegetable, fruit objects considering support vector machine (SVM) and k-nearest neighbor (kNN) classifiers using 3D color histogram features considered in this paper, for various values of number of color bins is shown in Table 5 and Table 6 respectively.

Table 5. Classification accuracies of various approaches on ALOI vegetable objects

Approach	3 bins	4 bins	5 bins	6 bins
SVM	46.74%	60.37%	66.81%	72.07%
kNN	79.85%	85.11%	89.18%	91.18%
CNN	95.7%	98.8%	99.26%	99.26%

Table 6. Classification accuracies of various approaches on ALOI fruit objects

Approach	3 bins	4 bins	5 bins	6 bins
SVM	47.41%	58.56%	64.01%	63.54%
kNN	80.74%	85.39%	87.51%	90.50%
CNN	96.3%	98.15%	98.41%	98.53%

From Table 5 and Table 6, it can be observed that accuracy improves with the increase in the number of color bins. Experiments conducted on varying number of color bins and different configurations of CNN suggest that the CNN classifier was able to recognize the objects from the 3D color histogram represented in 2D.

The high classification accuracy indicates that objects captured under varying illumination-color are also well classified suggesting that the CNN classifier was able to capture the illumination variations of objects. In contrast to the existing approaches like [15] [16], we considered a class of ALOI objects to evaluate the performance of the proposed approach. The proposed approach could not be compared with the existing approaches due to the dissimilarity in the objects considered for evaluation.

4 Conclusion

A view and illumination independent object classifier, using features derived from 3D color histogram and CNN architecture for classification is presented. The low classification error on ALOI fruits and vegetables objects suggests that the CNN classifier was able to capture the view and illumination invariant characteristics of the objects. The limitation of this approach is its inability to discriminate objects with same color profile in images. The future work includes extending this model to use 3D CNN on the 3D representation of 3D color histogram and to include shape and texture features for classification.

References

1. Andreopoulos, A., Tsotsos, J.K.: 50 years of object recognition: Directions forward. Computer Vision and Image Understanding **117** (2013) 827–891
2. Andrea Albarelli, Filippo Bergamasco, Luca Rossi, S. Vascon and Andrea Torsello: A stable graph-based representation for object recognition through high-order matching. In: International Conference on Pattern Recognition (ICPR), IEEE (2012) 3341–3344
3. Chauhan, A., Lopes, L.S.: Manhattan-pyramid distance: A solution to an anomaly in pyramid matching by minimization. In: International Conference on Pattern Recognition (ICPR), IEEE (2012) 2668–2672
4. Ya Su, Yun Fu, Xinbo Gao and Qi Tian: Discriminant learning through multiple principal angles for visual recognition. IEEE Transactions on Image Processing **21** (2012) 1381–1390
5. Meng Wang, Yue Gao, Ke Lu and Yong Rui: View-based discriminative probabilistic modeling for 3d object retrieval and recognition. IEEE Transactions on Image Processing **22** (2013) 1395–1407
6. Elazary, L., Itti, L.: A bayesian model for efficient visual search and recognition. Vision Research **50** (2010) 1338–1352
7. Zhuolin Jiang, Zhe Lin and Larry S. Davis: Label consistent k-svd: Learning a discriminative dictionary for recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence **35** (2013) 2651–2664
8. Serhat S. Bucak, Rong Jin and Anil K. Jain: Multiple kernel learning for visual object recognition: A review. IEEE Transactions on Pattern Analysis and Machine Intelligence **PrePrints** (2013) 1
9. Yoshua Bengio, Aaron C. Courville and Pascal Vincent: Representation learning: A review and new perspectives. IEEE Transactions on Pattern Analysis and Machine Intelligence **35** (2013) 1798–1828

10. Alex Krizhevsky, Ilya Sutskever and Geoffrey E. Hinton: Imagenet classification with deep convolutional neural networks. In: Neural Information Processing Systems (NIPS). (2012) 1106–1114
11. Pierre Sermanet, David Eigen, Xiang Zhang, Michael Mathieu, Rob Fergus and Yann LeCun: Overfeat: Integrated recognition, localization and detection using convolutional networks. Computer Research Repository (CoRR) (2013)
12. Bengio, Y.: Learning deep architectures for ai. Foundation and Trends in Machine Learning **2** (2009) 1–127
13. Palm, R.B.: Prediction as a candidate for learning deep hierarchical models of data. Master's thesis, Technical University of Denmark, Asmussens Alle, Denmark (2012)
14. Geusebroek, J.M., Burghouts, G.J., Smeulders, A.W.M.: The amsterdam library of object images. International Journal of Computer Vision **61** (2005) 103–112
15. Albarelli, A., Bergamasco, F., Rossi, L., Vascon, S., Torsello, A.: A stable graph-based representation for object recognition through high-order matching. In: ICPR, IEEE (2012) 3341–3344
16. Smagghe, P., Buessler, J.L., Urban, J.P.: Dj vu object localization using irf neural networks properties. In: IJCNN, IEEE (2013) 1–8