Unsupervised Learning Spatio-temporal Features for Human Activity Recognition from RGB-D Video Data

Guang Chen[1], Feihu Zhang[1], Manuel Giuliani[2], Christian Buckl[2], and Alois Knoll[1]

[1] Institut für Informatik VI, Technische Universität München, Boltzmannstr. 3, 85748 Garching, Germany
guang,zhang,knoll@in.tum.de
[2] fortiss GmbH, Guerickestr. 25, 80805 Munich, Germany
giuliani,buckl@fortiss.org

Abstract. Being able to recognize human activities is essential for several applications, including social robotics. The recently developed commodity depth sensors open up new possibilities of dealing with this problem. Existing techniques extract hand-tuned features, such as HOG3D or STIP, from video data. They are not adapting easily to new modalities. In addition, as the depth video data is low quality due to the noise, we face a problem: does the depth video data provide extra information for activity recognition? To address this issue, we propose to use an unsupervised learning approach generally adapted to RGB and depth video data. we further employ the multi kernel learning (MKL) classifier to take into account the combinations of different modalities. We show that the low-quality depth video is discriminative for activity recognition. We also demonstrate that our approach achieves superior performance to the state-of-the-art approaches on two challenging RGB-D activity recognition datasets.

Keywords: activity recognition, unsupervised learning, depth video.

1 Introduction

Human action recognition has been widely studied in computer vision. Its applications include video surveillance, content-based video search, robotics and a variety of systems that involve interactions between persons and computers. Traditional research mainly concentrates on learning and recognizing human activities from video data captured by a single visible light camera. The video data is a sequence of 2D frames with RGB or gray channels. There is extensive literature on action recognition for such video. Most methods for activity recognition with RGB video use hand-designed features like STIP [1], or use the local features like HOF [2] or HOG [3] to represent the spatio-temporal pattern. In these methods, human activities can be interpreted by a set of interesting points. However, there is no universally best hand-designed features for different RGB video [4]. In addition, it is difficult and time-consuming to extend these features

G. Herrmann et al. (Eds.): ICSR 2013, LNAI 8239, pp. 341–350, 2013.

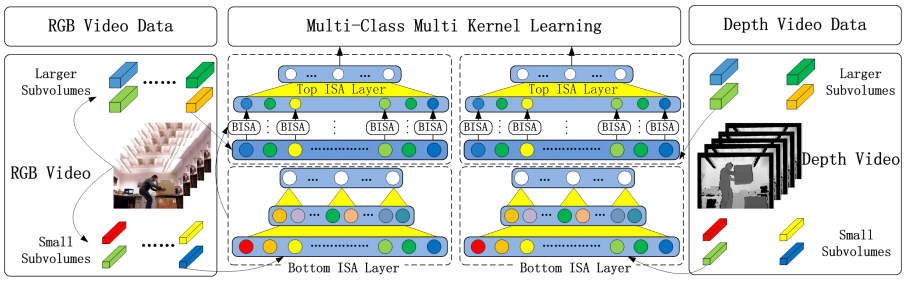


Fig. 1. An overview of our model: We randomly sample the subvolumes from RGB and depth video data. The small subvolumes are given as input to the Bottom ISA network. The learned bottom ISA model are copied to the Top ISA network. The stacked ISA learn the final features for each modality. The multi-class multi kernel learning is employed to learn a combination of different modalities and classify the activities (best viewed in color).

to other sensor modalities, such as laser scans or depth cameras. The depth camera can record RGB and depth video data has now become affordable and could be combined with standard vision system in social robot. The depth modality provides useful extra information to the complex problem of activity recognition since depth information is invariant to lighting and color variations.However, because there is no texture in the depth data,the extending hand-designed features from RGB data to depth modality are not discriminative enough for classifications. In addition, the depth video data is full of noises. There are large shadows or holes in the depth data. Hence the discrimination of the low quality depth video is considered doubtful.

Recently, there is a growing interest in unsupervised feature learning methods such as Deep Belief Nets [5], Sparse Coding [6, 7], Stacked Autoencoders [8], Independent Component Analysis (ICA) and Independent Subspace Analysis (ISA) [9]. These biologically-inspired learning algorithms show promise in the domain of the computer vision, such as object recognition with RGB-D images and action recognition with RGB video data [10–12]. Although many deep learning methods exist for learning features from RGB image or video data, none has yet been investigated for depth video data. In this paper, we provide an unsupervised learning method inspired by [9, 11]. We learn spatio-temporal features directly from RGB and depth video data independently. Fig. 1 outlines our approach. Our model starts with raw RGB and depth video data and extracts unlabeled space-time subvolumes from each modality. The subvolumes are then given to the stacked ISA network to learn hierarchical representations for each modality. We employ the multi kernel learning to combine the learned representations of different modalities. Our experiments show that although the low quality depth video data is full of noises, we could learn discriminative spatio-temporal features from depth data for activity recognition. We also achieve superior performance on the task of human activity recognition using RGB-D video data. Compared to other recent activity recognition methods for RGB-D video data [13–17], our ap-

proach is generalizable, does not need additional input channels such as skeleton joints and surface normals.

In this paper, we first briefly describe the ISA algorithm. Next we give details of how deep learning techniques such as convolution and stacking can be used to obtain hierarchical representations of the different modalities. Then, we learn the combinations of different modalities by multi kernel learning. The proposed features and models are evaluated on two RGB-D benchmark datasets: Pioneer-Activity dataset [13] and UTKinectAction3D dataset [14]. In our experiments, we show quantitative comparisons of different methods and different modalities.

2 Independent Subspace Analysis

In this section we describe the background of ISA algorithm [9]. ISA is an unsupervised learning algorithm that learns features from unlabeled subvolumes. First, random subvolumes are extracted into two sets, one for each modality(RGB and depth video data). Each set of subvolumes is then normalized and whitened. The pre-processed subvolumes are feed to ISA networks as input units. An ISA network [9] is described as a two-layer neural network, with square and square-root nonlinearities in the first and second layers respectively (see Fig. 1).

We start with any input unit $x^t \in \mathbb{R}^n$ for each random sampled subvolumes. We split each subvolume into a sequence of image patches and flatten them into a vector x^t with the dimension n. The activation of each second layer unit is

$$p_i(x^t; W, V) = \sqrt{\sum_{k=1}^{m} V_{ik} \left(\sum_{j=1}^{n} W_{kj} x_j^t\right)^2} \tag{1}$$

ISA learns parameters W through finding sparse feature representations in the second layer by solving

$$\begin{aligned} &\min_{W} \sum_{t=1}^{T} \sum_{i=1}^{m} p_i(x^t; W, V) \\ &s.t. WW^T = \mathbf{I} \end{aligned} \tag{2}$$

Here, $W \in \mathbb{R}^{k \times n}$ is the weights connecting the input units to the first layer units. $V \in \mathbb{R}^{m \times k}$ is the weights connecting the first layer units to the second layer units; n, k, m are the input dimension, number of the first layer units and second layer units respectively. The orthonormal constraint is to ensure the features are diverse.

The model so far has been unsupervised. The learned ISA filters for each modality could be used for activity recognitions. The first layer of our ISA model learns spatio-temporal features that detect a moving edge in time as shown in Fig. 2. It shows that the learned feature (each row) is able to group similar features in a group thereby achieving spatial invariance. When the method is applied to depth video data, the resulting filters have shaper edges which arise due to the strong discontinuities at object boundaries. The experiments in section 5.3 show that this property may contribute to a better recognition within depth video data compared to RGB video data. We also study the sensitivity of the

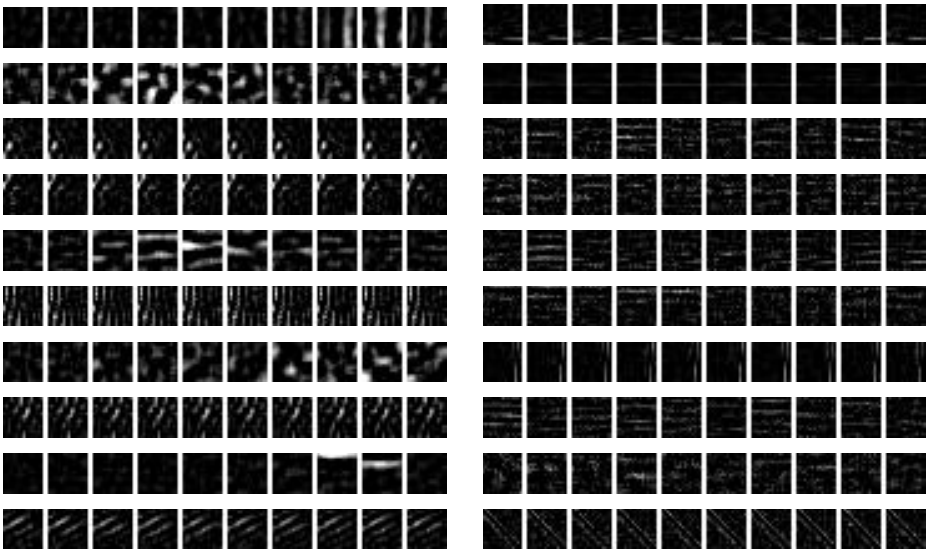


Fig. 2. Visualization of 20 ISA filters learned from PioneerActivity dataset. 10 filters (**left**) are from RGB video data and 10 filters (**right**) are from depth video data. These filters capture a moving edge in time. The filters from the depth video have sharper edges compared to filters trained on the RGB video data.

learned features to motion and orientation. In a control case, we limit this ability by using a temporal size of 4 frames instead of 10 frames and the recognition rate drops by 7.33% for the PioneerActivity dataset. If the temporal size is set to 2, the recognition rate drops by 2.56% again.

3 Stacked Convolutional ISA

3.1 Convolution and Stacking

In order to scale up to ISA algorithm to large input, we use a convolutional neural network architecture similar to [11, 18] for each modality. The network progressively makes use of PCA and ISA as sub-units for unsupervised learning as shown in Fig. 1.

We train the first layer of the networks with standard ISA algorithm on small input subvolumes for each modality. We randomly extract larger subvolumes from each modality. We then copy the learned bottom ISA filters and convolve with the larger subvolumes of the input video data (see Fig. 3). The responses of the convolution step are given as the input unit to the next ISA layer. As is common in neural network, we stack another ISA layer with PCA on top of the bottom ISA. We use PCA to whiten the data and reduce the dimensions of the input unit. The model is trained greedily layerwise in the same manner as other algorithms described in [5, 11, 19].

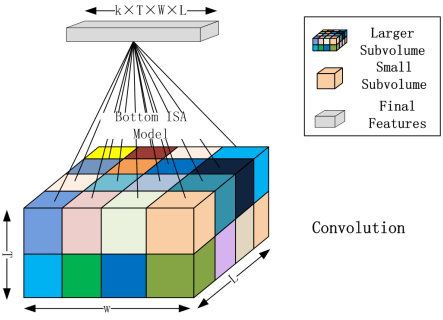


Fig. 3. Convolutional step of stacked ISA network. For clarity, the convolutional step is shown here non-overlapping, but in the experiments, convolution is done with overlapping.

3.2 Vector Quantization

As each activity is represented by a RGB video and an depth video, we perform the vector quantization by clustering the spatio-temporal features for each modality. We follow the state-of-the-art bag-of-words(Bow) paradigm. We construct the BoW features based on the dense spatio-temporal features for each modality.

4 Learning Multi-modality Combination

For each modality represented by the features of stacked convolutional ISA model, an SVM model on it defines a joint feature map $\Phi(x, y)$ on data $\mathcal{X}$ and labels $\mathcal{Y}$ as a linear output function $f_k(x, y) = \langle \omega_k, \Phi(x, y) \rangle + b_k$, parameterized with the hyperplane normal ω_k and bias b_k. The predicted class y for x is chosen to maximize the output $f_k(x, y)$.

Multi kernel learning considers a convex combination of n kernels, $K(x_i, x_j) = \sum_{k=1}^{n} \alpha_k K_k(x_i, x_j)$ where each kernel corresponds to a modality. We consider the following output function

$$f_{com}(x, y) = \sum_{k=1}^{n} [\alpha_k \langle \omega_k, \Phi(x, y) \rangle + b_k] \tag{3}$$

MKL learns the coefficient α, the weight ω and the bias b. For a multi class problem, different α, ω and b are learned for each class. In our case, we choose one-against-rest to decompose a multi-class problem. As MKL can not give a posterior class probability $P(y = 1|x)$, we propose approximating the posterior by a sigmoid function

$$P_m(y = 1|x) \approx P_{A_m, B_m}(f_{com}) \equiv \frac{1}{1 + \exp(A_m f_{com} + B_m)} \tag{4}$$

Fig. 4. Some example frames of two datasets. Samples in the top row are from the PioneerActivity dataset and samples in the bottom row are from the UTKinectAction3D dataset.

We follow Platt' method to learn A_m and B_m [20]. For each MKL-SVM model m, we learn a sigmoid function $P_{A_m,B_m}(f_{com})$. The maximum probability $l = \max_m P_m(y = 1|x)$ corresponds to the predicted label of x.

5 Experimental Results

We choose PioneerActivity dataset [13] and UTKinectAction3D dataset [14] to evaluate the proposed human activity recognition approach. Both datasets include the RGB video data and depth video data. The empirical results show the low-quality depth data provides more useful information for the task of human activity recognition. The results also show the proposed framework outperforms the state of art methods. We first give a brief overview of the datasets, followed by the detail of our stacked ISA model, and the experimental results.

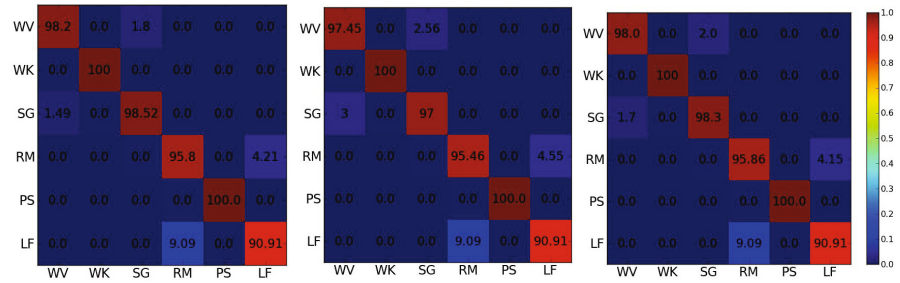
5.1 Datasets and Experimental Setup

The PioneerActivity dataset is an human activity dataset of RGB and depth video data captured by a depth camera [13]. The dataset presents several challenges due to illumination change, dynamic background and variations in human motions. It contains six types of human activities: lifting (LF), removing (RM), pushing (PS), waving (WV), walking (WK), signaling (SG). Each activity has 33 samples for each modality. We follow the experimental setup of [13]. We divide the database into three groups. One group as the training set, and the remaining groups are used as the testing sets. The experiment results are reported as the average over 20 runs.

The UTKinectAction3D dataset contains 10 types of human activities in indoor settings. The 10 actions include: sit down, stand up, walk, pick up, carry, throw, push, pull, wave and clap hands. Each action was collected from 10 different persons for 2 times. We evaluate our approach on this dataset using leaving one out cross validation. We run the experiment 20 times. Some samples of activities from the two datasets are shown in Fig. 4.

5.2 Model Details

We focus on the problem of human activity recognition from RGB and depth video data. We train stacked ISA model for each modality. For the PioneerActivity dataset, the input units to the bottom layer are of size $16 \times 16 \times 10$, 16, 16, 10 means spatial and temporal size of the subvolumes. The larger subvolumns to the top layer are of size $20 \times 20 \times 14$. The model parameters for the RGB and depth video data are the same. For the UTKinectAction3D dataset, the subvolumes to the bottom layer of stacked ISA network are of size $16 \times 16 \times 6$. The larger subvolumes to the top layer are of the size $20 \times 20 \times 8$. Finally, we performs vector quantization by K-means on the learned spatio-temporal features and classifies by multi-class MKL classifiers by χ^2 kernel.



(a) OursRGB-D (b) OursRGB (c) OursDepth

Fig. 5. The confusion matrices for the proposed method on PioneerActivity dataset with different modalities. Rows represent the actual classes, and columns represent predicted classes. OursRGB-D, OursRGB, OursD are our proposed method using both of RGB and depth data, using RGB data only, and using depth data only respectively. (best viewed in color).

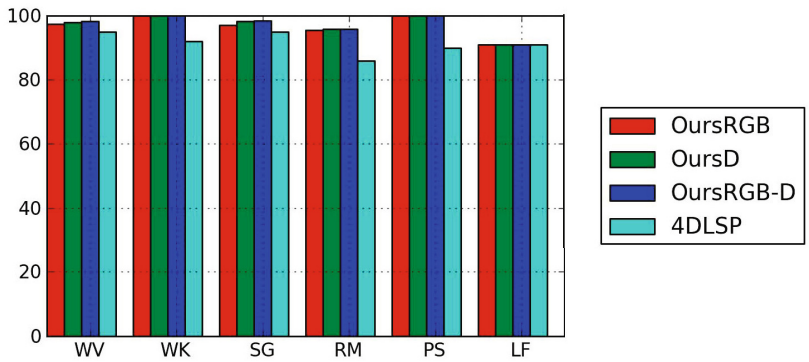


Fig. 6. The comparison between the average accuracy of the proposed method and 4DLSP [13] on the PioneerActivity dataset with different modalities

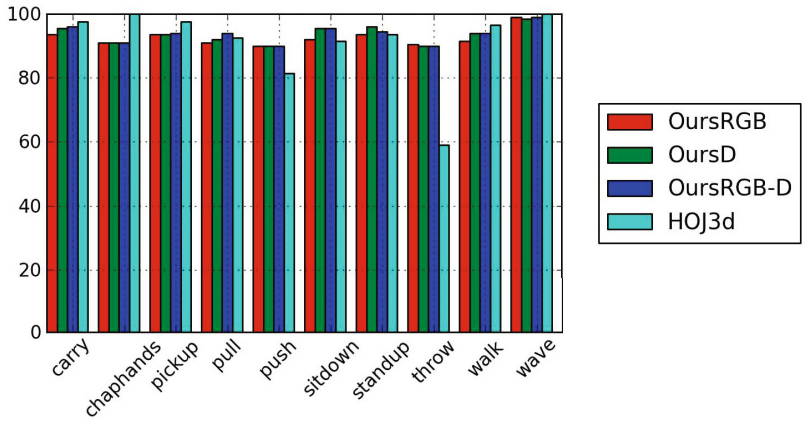


Fig. 7. The comparision between the average accuracy of the proposed method and HOJ3d [14] on the UTKinectAction dataset with different modalities

Table 1. Summary of the experimental results for PioneerActivity dataset and UTKinectAction3D dataset

Ave. acc on two Datasets	OursRGB	OursD	OursRGB-D	4DLSP [13]	HOJ3d [14]
PioneerActivity dataset	96.80	97.17	97.23	91.5	-
UTKinectAction3D dataset	92.55	93.6	93.8	-	90.95

5.3 Experimental Results

We shows the accuracies of our method on the PioneerActivity dataset in Fig.5 and Fig 6. The confusion matrices of our approach using different modalities are given in Fig. 5. The confusion matrix shows that the largest confusion lies between "removing"and "lifting". This is consistent with the bag-of-word paradigm as it assumes each word is independent of others. In Table 1, we compare our test results with the state of the art method [13]. Our method significantly outperforms 4DLSP [13].

We further compare our approach with the best published result on UTKinectAction3D dataset [14] in Fig. 7 and Table 1. The average accuracy of our method using RGB-D video data is 93.8%. Notice that the work of [14] only uses the depth video data for activity recognition. Our approach with the depth video data achieves 93.6% accuracy which is still better than HOJ3D [14].

5.4 Discussion

In the above experiments, we compared our approach using the combination of the RGB video data and depth video data. This raises some questions: "How much does the combination help?"and "Does the depth video data provide extra information for activity recognition? ". In Table 1, the accuracies of the methods

using both modalities are just little better than the approach using only one modality. This result shows that the learned features from different modalities exhibit some similar patterns. One possible explanation would be the features learned from RGB video data is trained on gray scale versions of the RGB video data.

In general, the accuracies of the approach using depth video data are better than the accuracies of the method using RGB video data. Although the depth data is full of noise, such as shadows and holes, the result indicates that the depth video data provides more useful information for the task of human activity recognition than the RGB video data on our datasets. A possible explanation is that RBG video data is sensitive to the illumination variations and the dynamics background while the depth video data is not. As illustrated in Fig. 4, the computer monitors lead to a dynamic background for the RGB images, which distract the learned features from capturing useful human motions. But the depth sensor is not sensitive to the dynamic background. Another explanation would be the learned features by the depth video data capture more edge informations. Because the edge detectors like Gabor filters show great performance in a lot of computer vision domains [21]. This is consistent with Fig.2 that the resulting filters learned by depth data have shaper edges.

6 Conclusion

We introduced a method that learns spatio-temporal features from RGB-D video data. The stacked ISA network learns the hierarchical representations in an unsupervised way. The multi-class MKL learns the combinations of different modalities. This architecture could leverage the plethora of the unlabeled data and adapt easily to new modalities. The experiment results were carried out with PioneerActivity and UTKinectAction datasets. We observed that the low-quality depth video data provides more useful information for the task of human activity recognition on our datasets. The learned features from RGB and depth video data exhibit some similar patterns. Our method also outperforms the state-of-the-art methods for activity recognition.

References

1. Laptev, I.: On space-time interest points. Int. J. Comput. Vision 64(2-3), 107–123 (2005)
2. Laptev, I., Marszałek, M., Schmid, C., Rozenfeld, B.: Learning realistic human actions from movies. In: Conference on Computer Vision & Pattern Recognition (June 2008)
3. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: CVPR, pp. 886–893 (2005)
4. Wang, H., Ullah, M.M., Kläser, A., Laptev, I., Schmid, C.: Evaluation of local spatio-temporal features for action recognition. In: British Machine Vision Conference, p. 127 (September 2009)

5. Hinton, G.E., Osindero, S., Teh, Y.W.: A fast learning algorithm for deep belief nets. Neural Computation 18(7), 1527–1554 (2006)
6. Olshausen, B.A., Field, D.J.: Emergence of simple-cell receptive field properties by learning a sparse code for natural images. Nature 381, 607–609 (1996)
7. Lee, H., Battle, A., Raina, R., Ng, A.Y.: Efficient sparse coding algorithms. In: NIPS, pp. 801–808 (2007)
8. Bengio, Y., Lamblin, P., Popovici, D., Larochelle, H.: Greedy layer-wise training of deep networks. In: Schölkopf, B., Platt, J., Hoffman, T. (eds.) Advances in Neural Information Processing Systems 19, pp. 153–160. MIT Press, Cambridge (2007)
9. Hyvrinen, A., Hurri, J., Hoyer, P.O.: Natural Image Statistics: A Probabilistic Approach to Early Computational Vision, 1st edn. Springer Publishing Company, Incorporated (2009)
10. Socher, R., Huval, B., Bath, B.P., Manning, C.D., Ng, A.Y.: Convolutional-recursive deep learning for 3d object classification. In: Bartlett, P.L., Pereira, F.C.N., Burges, C.J.C., Bottou, L., Weinberger, K.Q. (eds.) NIPS, pp. 665–673 (2012)
11. Le, Q., Zou, W., Yeung, S., Ng, A.: Learning hierarchical invariant spatio-temporal features for action recognition with independent subspace analysis. In: 2011 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3361–3368 (2011)
12. Ji, S., Xu, W., Yang, M., Yu, K.: 3d convolutional neural networks for human action recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence 35(1), 221–231 (2013)
13. Zhang, H., Parker, L.: 4-dimensional local spatio-temporal features for human activity recognition. In: 2011 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pp. 2044–2049 (2011)
14. Xia, L., Chen, C.C., Aggarwal, J.: View invariant human action recognition using histograms of 3d joints. In: 2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 20–27 (2012)
15. Yang, X., Tian, Y.: Eigenjoints-based action recognition using nave-bayes-nearest-neighbor. In: CVPR Workshops, pp. 14–19. IEEE (2012)
16. Wang, J., Liu, Z., Wu, Y., Yuan, J.: Mining actionlet ensemble for action recognition with depth cameras. In: 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1290–1297 (2012)
17. Oreifej, O., Liu, Z.: Hon4d: Histogram of oriented 4d normals for activity recognition from depth sequences (June 2013)
18. LeCun, Y., Boser, B., Denker, J.S., Henderson, D., Howard, R.E., Hubbard, W., Jackel, L.D.: Backpropagation applied to handwritten zip code recognition. Neural Computation 1, 541–551 (1989)
19. Bengio, Y., Lamblin, P., Popovici, D., Larochelle, H.: Greedy layer-wise training of deep networks, pp. 153–160 (2007)
20. Platt, J.C.: Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In: Advances in Large Margin Classifier, pp. 61–74. MIT Press (1999)
21. Kamarainen, J.K., Kyrki, V., Kälviäinen, H.: Invariance properties of Gabor filter based features - overview and applications. IEEE Transactions on Image Processing 15(5), 1088–1099 (2006)