Multimodal Sparse Features for Object Detection

Martin Haker, Thomas Martinetz, and Erhardt Barth

Institute for Neuro- and Bioinformatics, University of Lübeck,
Ratzeburger Allee 160, 23538 Lübeck, Germany
{haker,martinetz,barth}@inb.uni-luebeck.de
http://www.inb.uni-luebeck.de

Abstract. In this paper the sparse coding principle is employed for the representation of multimodal image data, i.e. image intensity and range. We estimate an image basis for frontal face images taken with a Time-of-Flight (TOF) camera to obtain a sparse representation of facial features, such as the nose. These features are then evaluated in an object detection scenario where we estimate the position of the nose by template matching and a subsequent application of appropriate thresholds that are estimated from a labeled training set. The main contribution of this work is to show that the templates can be learned simultaneously on both intensity and range data based on the sparse coding principle, and that these multimodal templates significantly outperform templates generated by averaging over a set of aligned image patches containing the facial feature of interest as well as multimodal templates computed via Principal Component Analysis (PCA). The system achieves a detection rate of 96.4% on average with a false positive rate of 3.7%.

1 Introduction

In recent years there has been a lot of interest in learning sparse codes for data representation, and favorable properties of sparse codes with respect to noise resistance have been investigated [1]. Olshausen and Field [2] applied sparse coding to natural images and showed that the resulting features resemble receptive fields of simple cells in V1. Thus, it stands to reason that the basis functions computed by sparse coding can be used effectively in pattern recognition tasks in the fashion introduced by Serre et al. [3], who model a recognition system that uses cortex-like mechanisms.

Sparse coding has also been successfully applied to the recognition of handwritten digits [4]. The authors learn basis functions for representing patches of handwritten digits and use these to extract local features for classification.

In this work, we aim to learn a sparse code for multimodal image data, i.e. we simultaneously learn basis functions for representing corresponding intensity and range image patches. As a result, we obtain aligned pairs of basis functions that encode prominent features that co-occur consistently in both types of data. Thus, a corresponding pair of basis functions can be used to consistently extract

C. Alippi et al. (Eds.): ICANN 2009, Part II, LNCS 5769, pp. 923–932, 2009.

features from intensity and range data. To our knowledge, sparse representations have not yet been learned for multimodal signals.

The considered image data was obtained by a Time-of-Flight (TOF) camera [5] which provides a range map that is perfectly registered with an intensity image (often referred to as an *amplitude* image in TOF nomenclature). Although TOF cameras emerged on the market only recently, they have been used in a number of image processing applications, such as shape from shading [6], people tracking [7], gesture recognition [8], and stereo vision [9]. A review of publications related to TOF cameras can be found in [10].

It has already been shown that using both intensity and range data of a TOF camera in an object detection task can significantly improve performance in comparison to using either data alone [11]. The fact, that a sparse code learned simultaneously on both intensity and range data yields perfectly aligned basis functions, allows us to extract relevant features from both types of data.

Here, we aim to learn a set of basis functions that encode structural information of frontal face images in a component-based fashion. As a result, the basis functions estimated by sparse coding can be regarded as templates for facial features, such as the nose. We evaluate the resulting templates on a database of TOF images and use simple template matching to identify the presence and position of the nose in frontal face images. The importance of the nose as a facial feature for problems such as head tracking was already mentioned in [12,13].

Section 2 will discuss the computation of a set of basis functions under the constraint of the sparse coding principle. In Section 3 we discuss the procedure of determining the basis function that yields the optimal equal error rate (EER) in the nose detection task. Section 4 presents the results and shows that templates generated via sparse coding yield significantly better detection rates than templates obtained by PCA or by averaging over a set of aligned image patches.

2 Sparse Features

The investigated database of frontal face images [14] was obtained using an SR3000 TOF camera [15]. The subjects were seated at a distance of about 60 cm from the camera and were facing the camera with a maximum horizontal and/or vertical head rotation of approximately 10 degrees. As a result, the facial feature of interest, i.e. the nose, appears at a size of roughly 10×10 pixels in the image. A number of sample images are given in Fig. 1.

As a TOF camera provides a range map that is perfectly registered with an intensity image, we aim to learn an image basis for intensity and range simultaneously. To this end, the input data for the sparse coding algorithm are vectors whose first half is composed of intensity data and the second half of range data, i.e. in case we consider image patches of size 13×13, each patch is represented by a 338-dimensional vector ($d = 338 = 2 \cdot 13 \cdot 13$) where the first 169 dimensions encode intensity and the remaining 169 dimensions encode range.

In order to speed up the training process, we only considered training data that originated from an area of 40×40 pixels centered around the position

Fig. 1. Three sample images of frontal face images taken by an SR3000 TOF camera. The top row shows the intensity and the bottom row the range data.

of the nose. The position of the nose was annotated manually beforehand. By this procedure we prevent the basis functions from being attracted by irrelevant image features, and a number of 72 basis functions proved to be sufficient to represent the dominant facial features, such as the nose or the eyes.

A common difficulty with TOF images is that the range data is relatively noisy and that both intensity and range can contain large outliers due to reflections of the active illumination (e.g. if subjects wear glasses). These outliers violate the assumed level of Gaussian additive noise in the data and can lead the sparse coding algorithm astray. To compensate for this effect, we applied a 5×5 median filter to both types of data. To ensure the conservation of detailed image information while effectively removing only outliers, pixel values in the original image I_o were only substituted by values of the median filtered image I_f if the absolute difference between the values exceeded a certain threshold:

$$I_o(i,j) = \begin{cases} I_f(i,j) & \text{if } |I_o(i,j) - I_f(i,j)| \geq \theta \\ I_o(i,j) & \text{otherwise} \end{cases} .$$

There exist a number of different sparse coding approaches, see for example [2,16,17]. We employed the Sparsenet algorithm [2] for learning the sparse code. The basic principle aims at finding a basis W for representing vectors $\boldsymbol{x}$ as a linear combination of the basis vectors using coefficients $\boldsymbol{a}$ under the assumption of Gaussian additive noise: $\boldsymbol{x} = W\boldsymbol{a} + \boldsymbol{\epsilon}$. To enforce sparseness, i.e. the property that the majority of coefficients a_i are zero, the Sparsenet algorithm solves the following optimization problem:

$$\min_W E\left(\min_{\boldsymbol{a}} \left(\|\boldsymbol{x} - W\boldsymbol{a}\| + \lambda S(\boldsymbol{a})\right)\right) . \tag{1}$$

Here, E denotes the expectation and $S(\boldsymbol{a})$ is an additive regularization term that favors model parameters W that lead to sparse coefficients $\boldsymbol{a}$. The parameter λ balances the reconstruction error $\boldsymbol{\epsilon}$ against the sparseness of the coefficients.

In order to apply the method, the input data has to be whitened beforehand as indicated in [2]. We applied the whitening to both types of data individually. Only after this preprocessing step, the training data was generated by selecting random image patches of the template size, i.e. for a patch in a given image the corresponding intensity and range data were assembled in a single vector.

The resulting features for 19×19 image patches are given in Fig. 2. Facial features, e.g. nose, eyes, and mouth, can readily be distinguished. We set the parameter λ to a relatively high value ($\lambda = 0.1$), i.e. we enforce high sparseness, in order to obtain this component-based representation, however we can report that the results are not particularly sensitive to minor changes of this parameter.

3 Nose Detection

Since the basis functions computed by sparse coding in Section 2 represent facial features, it stands to reason that they can be used for object detection via template matching. At this point two questions arise: (i) Which basis function represents the best template, and (ii) what is the actual position of the facial feature with respect to the center of the image patch corresponding to this basis function. A straightforward solution would be to select the most promising feature by visual inspection and to annotate the position of the facial feature within the image patch manually. Obviously though, this procedure is not generally applicable and is likely to yield suboptimal results.

Thus, we decided to follow a computationally more intensive procedure that, in contrast, is fully automatic and operates in a purely data-driven fashion. For each of the 72 basis functions we trained and evaluated a nose detector for every possible position of the nose in a certain neighborhood around the center of the image patch. In the case of 13×13 image patches we chose this neighborhood to be 11×11. As a result, a total of $8712 = 72 \cdot 11 \cdot 11$ detectors were trained. The final detector uses the basis function and the position of the nose out of the 8712 configurations that produced the best EER on the training set.

The thresholds of each detector were simply determined by taking the minimum and the maximum of the filter responses at the annotated positions of the nose on a set of training images, i.e. upper and lower bounds for the filter responses that identify a nose were determined for both intensity and range data. In order to identify a nose in a new image, both intensity and range were filtered with the corresponding template images and each pixel whose filter responses complied with the identified bounds was classified as a nose pixel. To obtain an EER, these bounds were relaxed or tightened.

The procedure was evaluated on a data set of 120 TOF images of frontal faces taken from 15 different subjects. To double the amount of data, mirrored versions of the images were also added to the data set. From the total of 240 images one half was chosen at random as a training set to determine the bounds of each classifier. These bounds were then adjusted to yield an EER on the training set. Finally, the optimal classifier, i.e. the one out of the 8712 candidates yielding

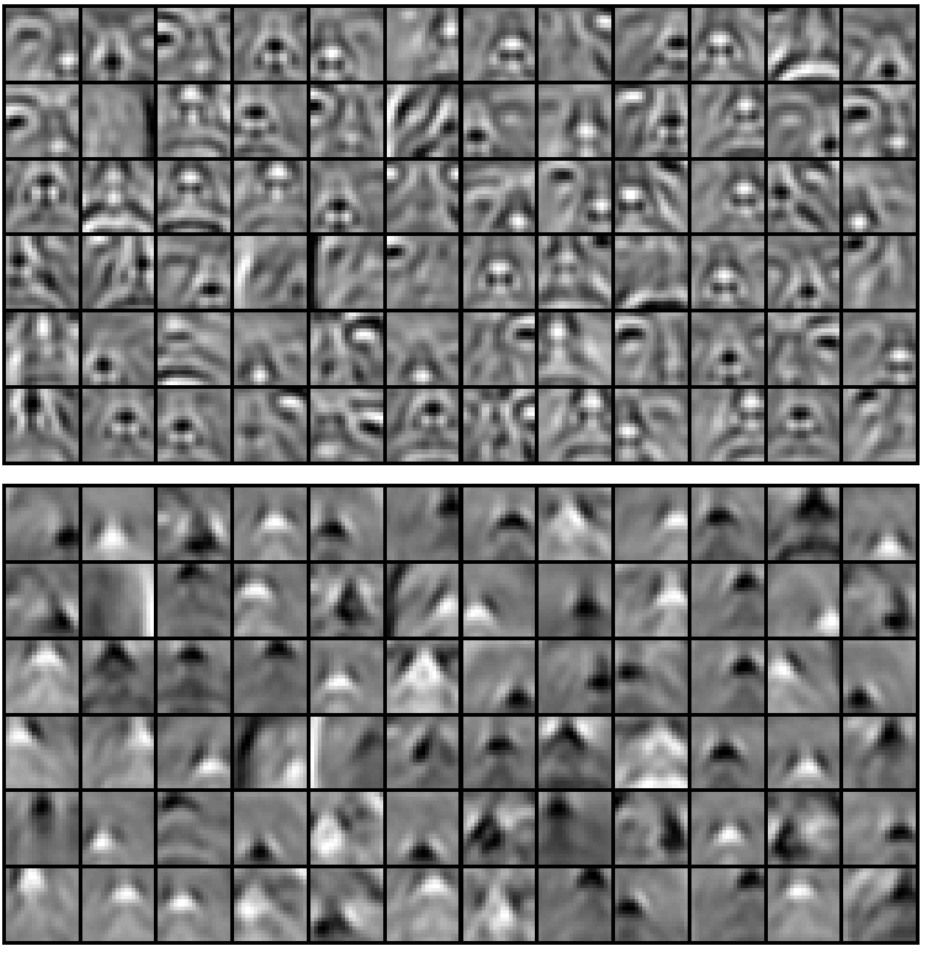

Fig. 2. Basis functions learned for frontal face images via the Sparsenet algorithm. The upper and lower part of the figure show the basis functions representing intensity data and range data, respectively. The basis functions for both types of data were learned simultaneously and correspond pairwise, i.e. the top-left intensity feature is perfectly aligned with the top-left range feature.

the best EER, was evaluated on the remaining 120 images that were not used during the training process.

In order to assess the performance of the learned templates, we also evaluated two other types of templates – "average" and "eigen" templates. The former were generated by averaging over a set of aligned image patches containing a nose. The latter were obtained as the principal components of these aligned image patches. Again, we generated corresponding pairs of templates for both

intensity and range data. The same training images, including the preprocessing, were used as in Section 2 for the Sparsenet algorithm.

A fundamental difference between these two approaches to generating the average and eigen templates and the sparse coding method is, that the former only yield templates in which the nose is centered in the image patch whereas the latter also produces translated versions of the nose (see Fig. 2). To guarantee a fair comparison between the different templates we applied the following procedure: since the optimal position of the nose within the template is not known a priori, we generated a total of 121 13×13 templates centered at all possible positions in a 11×11 neighborhood around the true position of the nose, i.e. the templates were shifted so that the nose was not positioned in the center of the template. In correspondence to the procedure described above for the sparse-coding templates, each shifted template was then evaluated for every possible position of the nose in a 11×11 neighborhood around the center of the image patch. For the average templates the resulting number of possible detectors amounts to 14641. In the case of the eigen templates, it is not apparent which principal component should be used as a template. To constrain the computational complexity, we considered only the first three principal components. Nevertheless, this resulted in 43923 possible detectors. Again, the optimal average and eigen templates were determined as the ones yielding the best EER on the training set according to the procedure described above.

4 Results

The results of the training for the nose detection task using 13×13 templates are given in Fig. 3. The EER on the training set using the sparse-coding templates is 3.9%. The eigen templates achieve an EER of 6.6%, and the average templates yield an EER of 22.5%, i.e. the EERs for these two procedures are higher by roughly 50% and 500%, respectively. The EERs prove to be largely independent of the training set. We ran 100 evaluations of the procedure with random configurations of the training set and recomputed both the templates and the classifiers in each run. The standard deviations for the three EERs over the 100 evaluations were $\sigma = 0.9\%$, $\sigma = 1.6\%$, and $\sigma = 2.3\%$, respectively.

Fig. 3 also shows the ROC curves for detectors that use the sparse-coding templates computed on either intensity or range data of the TOF images alone. Note that the EERs are dramatically higher in comparison to the detector that uses both types of data together. This confirms results reported in [11], where the combination of intensity and range data also yielded markedly better results in the detection of the nose based on geometrical features.

The error rates on the test set are only slightly higher than the EERs on the training set, which shows that the method generalizes well to new data. The false positive rates (FPR) amount to 5.3%, 9.3%, and 24.4% for the sparse-coding, eigen, and average templates.

The evaluation above considered only templates of a fixed size of 13×13 pixels. However, varying the template size reveals some interesting properties of

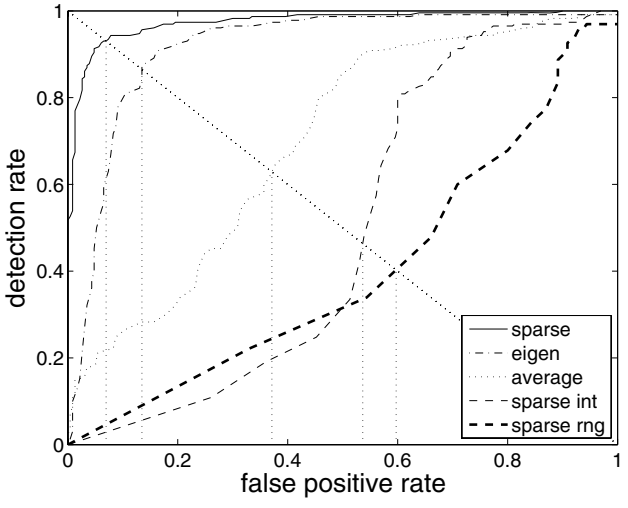

Fig. 3. ROC curves of detection rate vs. false positive rate. The curves were generated using the **sparse**-coding templates, the **eigen** templates, the **average** templates, and the sparse-coding templates using only the intensity data (**sparse int**) or only the range data (**sparse rng**). The detection rate gives the percentage of images in which the nose was identified correctly, whereas the false positive rate denotes the percentage of images where at least one non-nose pixel was misclassified. Thus, strictly speaking, the curves do not represent ROC curves in the standard format, but they convey exactly the information one is interested in for this application, that is, the accuracy with which the detector gives the correct response per image.

the different approaches. To this end, we computed templates of size $n \times n$, where $n = 1, 3, \ldots, 19$, for each approach and estimated the optimal detector according to the same procedure outlined above. To reduce the number of possible detectors to evaluate, the neighborhood sizes for positioning the nose and shifting the template were reduced to 7×7 pixels for templates with size n smaller than 13. Again, we considered 100 random configurations of training and test set.

Fig. 4 shows the configurations of template and position of the nose within the template that yielded the best EERs on the training set for each approach with respect to the different template sizes.

Note that the sparse-coding templates (first two rows) exhibit a much higher contrast in comparison to the average templates (rows three and four), especially for larger sizes of the template. This explains the bad performance of the average templates, because an increase in size does not add more information to the template. This effect becomes clearly visible in Fig. 5. The plot shows the FPRs on the test set for the different approaches over varying template sizes. One can observe that the FPR of the average template starts to increase for templates of size five. In comparison, the FPR of the sparse-coding templates continues to decrease up to a template size of 19.

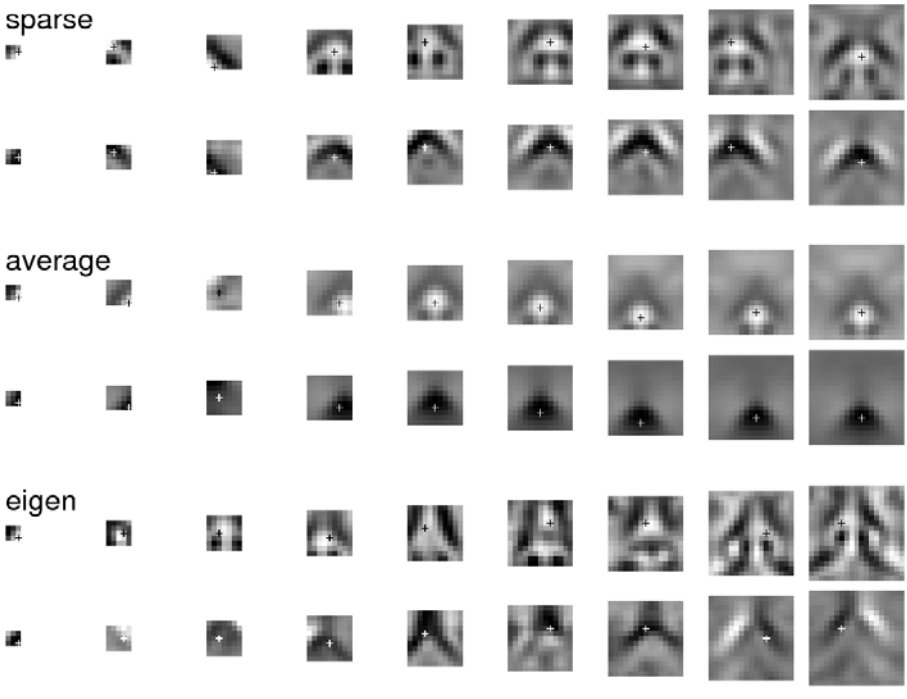

Fig. 4. Optimal templates for different template sizes, where each column shows templates of odd pixel sizes ranging from 3 to 19. The first row shows the sparse-coding templates for intensity data and the second row shows the corresponding features for range data. Rows three and four give the average templates and rows five and six show eigen templates. The crosses mark the estimated position of the nose within the templates.

A decrease of the FPR can also be observed for the eigen templates up to size 11 of the template. For larger template sizes the FPR also starts to increase, whereas the sparse-coding templates continue to achieve low FPRs, as already mentioned above. It seems that sparse coding can exploit further reaching dependencies.

A comparison of the FPRs with respect to the optimal template size for each method reveals that the average template achieves the worst overall performance with an FPR of 9.6% ($\sigma = 3.5$) for a 7×7 template. The best results for the eigen templates were obtained with templates of size 11 yielding an FPR of 7.9% ($\sigma = 3.2$). The sparse coding templates of size 19 had the best overall performance (FPR 3.7%, $\sigma = 2.3$), and the FPR improved roughly by a factor of 2.5 in comparison to the best eigen template.

Note that the false negative rate for the different approaches lies well within the error bars of the FPR in Fig. 5, as one would expect, since the classifier was set to achieve an EER during training.

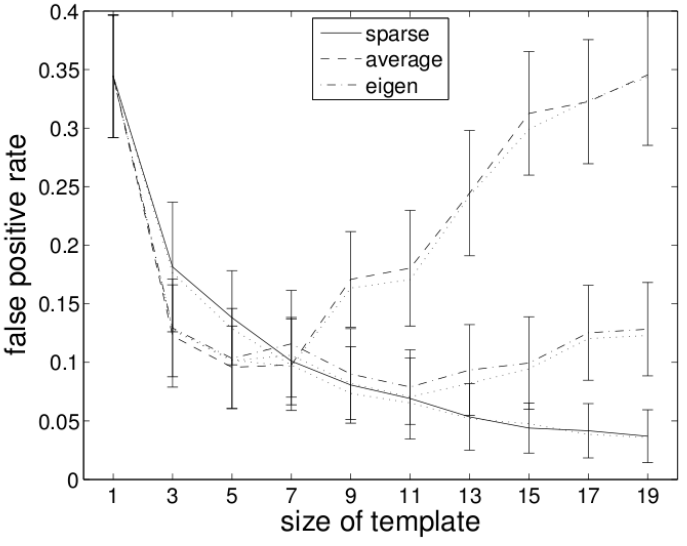

Fig. 5. The graph shows the FPRs and standard deviations on the different test sets for the different templates at different template sizes. The dotted lines show the corresponding false negative rates.

5 Discussion

We have demonstrated how a sparse code can be learned for multimodal image data. The resulting basis functions can be used effectively for template matching in detecting the nose in frontal face images. The sparse-coding templates yield significantly improved results in comparison to templates obtained by averaging over a number of aligned sample images of noses. Templates resembling the principal components of these aligned sample images were also outperformed, especially for large sizes of the template.

The sparse-coding templates were learned on intensity and range data of a TOF camera simultaneously, which yields templates that are perfectly registered for the two different input modalities. The combination of intensity and range data yields a greatly improved detector compared to either type of data alone.

Acknowledgment

This work was developed within the ARTTS project (www.artts.eu), which is funded by the European Commission (contract no. IST-34107) within the Information Society Technologies (IST) priority of the 6th Framework Programme. This publication reflects the views only of the authors, and the Commission cannot be held responsible for any use which may be made of the information contained therein.

References

1. Donoho, D.L., Elad, M., Temlyakov, V.N.: Stable recovery of sparse overcomplete representations in the presence of noise. IEEE Transactions on Information Theory 52(1), 6–18 (2006)
2. Olshausen, B.A., Field, D.J.: Sparse coding with an overcomplete basis set: A strategy employed by V1? Vision Research 37, 3311–3325 (1997)
3. Serre, T., Wolf, L., Bileschi, S., Riesenhuber, M., Poggio, T.: Robust object recognition with cortex-like mechanisms. IEEE Transactions on Pattern Analysis and Machine Intelligence 29(3), 411–426 (2007)
4. Labusch, K., Barth, E., Martinetz, T.: Simple Method for High-Performance Digit Recognition Based on Sparse Coding. IEEE Transactions on Neural Networks 19(11), 1985–1989 (2008)
5. Oggier, T., Büttgen, B., Lustenberger, F., Becker, G., Rüegg, B., Hodac, A.: SwissRanger^{TM} SR3000 and first experiences based on miniaturized 3D-TOF cameras. In: Ingensand, K. (ed.) Proc. 1^{st} Range Imaging Research Day, Zurich, pp. 97–108 (2005)
6. Böhme, M., Haker, M., Martinetz, T., Barth, E.: Shading constraint improves accuracy of time-of-flight measurements. In: CVPR 2008 Workshop on Time-of-Flight-based Computer Vision, TOF-CV (2008)
7. Hansen, D.W., Hansen, M., Kirschmeyer, M., Larsen, R., Silvestre, D.: Cluster tracking with time-of-flight cameras. In: CVPR 2008 Workshop on Time-of-Flight-based Computer Vision, TOF-CV (2008)
8. Kollorz, E., Penne, J., Hornegger, J., Barke, A.: Gesture recognition with a Time-Of-Flight camera. International Journal of Intelligent Systems Technologies and Applications 5(3/4), 334–343 (2008)
9. Gudmundsson, S.A., Aanaes, H., Larsen, R.: Fusion of Stereo Vision and Time-of-Flight Imaging for Improved 3D Estimation. In: Dynamic 3D Imaging – Workshop in Conjunction with DAGM (2007) (in print)
10. Kolb, A., Barth, E., Koch, R., Larsen, R.: Time-of-Flight Sensors in Computer Graphics. Eurographics State of the Art Reports, 119–134 (2009)
11. Haker, M., Böhme, M., Martinetz, T., Barth, E.: Geometric invariants for facial feature tracking with 3D TOF cameras. In: Proceedings of the IEEE International Symposium on Signals, Circuits & Systems (ISSCS), Iasi, Romania, vol. 1, pp. 109–112 (2007)
12. Yin, L., Basu, A.: Nose shape estimation and tracking for model-based coding. In: Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2001), May 2001, vol. 3, pp. 1477–1480 (2001)
13. Gorodnichy, D.O.: On importance of nose for face tracking. In: Proceedings of the IEEE International Conference on Automatic Face and Gesture Recognition (FG 2002), Washington, D.C, May 2002, pp. 188–196 (2002)
14. ARTTS: 3D TOF Database, `http://www.artts.eu/publications/3d_tof_db`
15. Oggier, T., Büttgen, B., Lustenberger, F., Becker, G., Rüegg, B., Hodac, A.: SwissRanger^{TM} SR3000 and first experiences based on miniaturized 3D-TOF cameras. In: Proceedings of the 1st Range Imaging Research Day, Zürich, Switzerland, pp. 97–108 (2005)
16. Lewicki, M.S., Sejnowski, T.J., Hughes, H.: Learning overcomplete representations. Neural Computation 12, 337–365 (2000)
17. Labusch, K., Barth, E., Martinetz, T.: Sparse Coding Neural Gas: Learning of Overcomplete Data Representations. Neurocomputing (2009) (in press)