Automatic Localization and Identification of Vertebrae in Arbitrary Field-of-View CT Scans

Ben Glocker[1], J. Feulner[2], Antonio Criminisi[1], D.R. Haynor[3], and E. Konukoglu[1]

[1] Microsoft Research, Cambridge, UK
[2] University of Erlangen-Nuremberg, Erlangen, Germany
[3] University of Washington, Seattle, WA, USA

Abstract. This paper presents a new method for automatic localization and identification of vertebrae in arbitrary field-of-view CT scans. No assumptions are made about which section of the spine is visible or to which extent. Thus, our approach is more general than previous work while being computationally efficient. Our algorithm is based on regression forests and probabilistic graphical models. The discriminative, regression part aims at roughly detecting the visible part of the spine. Accurate localization and identification of individual vertebrae is achieved through a generative model capturing spinal shape and appearance. The system is evaluated quantitatively on 200 CT scans, the largest dataset reported for this purpose. We obtain an overall median localization error of less than 6mm, with an identification rate of 81%.

1 Introduction

This paper proposes an algorithm for automatic detection, localization, and identification of individual vertebrae in computed tomography scans. A variety of tasks beyond spine specific analysis can immediately benefit from such a system. The spine provides a natural patient-specific coordinate system, where individual vertebrae serve as anatomical landmarks. These can be used, for instance, for semantically guided inspection tools, linking of radiological reports with corresponding image regions, or for robust initialization of image registration. Vertebrae localization also provides valuable priors for subsequent tasks such as anatomy segmentation, image retrieval, shape and population analysis.

The challenges associated with automatic localization of individual vertebrae arise from i) the repetitive nature of these structures, ii) the variability of normal and pathological anatomy, iii) and the variability of images (*e.g.* resolution and field-of-view). A common approach for vertebra (in CT) and intervertebral disks (in MRI) is to employ a multi-stage approach. In the first stage a detector in the form of a filter [1, 2], a single/multi-class classifier [3–8] or a model-based Hough transform [9] is used to detect potential vertebra candidates. As these candidates may contain many false positive responses a second stage is applied to add robustness. Prior knowledge on the global shape and/or appearance of individual vertebrae and their interconnections is used. In [2], a clever search is

N. Ayache et al. (Eds.): MICCAI 2012, Part III, LNCS 7512, pp. 590–598, 2012.
© Springer-Verlag Berlin Heidelberg 2012

performed based on prior information through the candidates, while [1, 4, 5] fit a low order polynomial curve to the candidates to remove outliers. In [3, 6, 8, 9] the authors add prior information via graphical models, such as Hidden Markov Models (HMMs) [10], and infer the maximum a-posteriori (MAP) estimate for the vertebrae locations. In contrast, [11, 12] use a fully generative model, and inference is achieved via generalized expectation-maximization, while in [7] deformable templates are used for segmentation and subsequent identification.

Although previous works achieve high localization accuracy they cannot handle completely general scans where it is not known in advance which portion of the anatomy is visible. In fact, most algorithms require a priori knowledge of which vertebrae are visible in the scan [1–8]. They either focus on a specific region, such as lumbar or thoracic, or need to modify their models based on the expected spine region. In [11, 12] approximate alignment between scans is assumed. To the best of our knowledge the only work that explicitly handles arbitrary scans is [9]. However, the added generality comes at high computational cost. Based on an affine vertbra registration algorithm, the identification phase is reported to take up to 36 minutes for 12 thoracic vertebrae.

In this paper we overcome those drawbacks with a vertebra localization and identification algorithm which is both robust and efficient. Its main advantage is the automatic handling of arbitrary field-of-view scans displaying widely varying anatomical regions. For instance, in a narrow abdominal scan we may be able to see just a handful of vertebrae together with the kidneys. A radiologist makes use of such contextual information to infer that we are looking at a lumbar section of the spine. In our system we incorporate contextual information within a regression forest algorithm. More specifically, we build upon state-of-the-art supervised, non-linear regression techniques [13] used jointly with a probabilistic, generative prior of spinal shape and appearance. The forest provides context-aware, fast estimation of vertebrae centres. In a second stage a joint model of vertebra appearance and global spine shape yields a refined localization as well as individual vertebra identification. The whole process takes less than 2 minutes on a standard desktop machine, thus allowing integration in existing image analysis pipelines. Details of our approach are presented in the next section, followed by an extensive quantitative validation on a large labelled dataset of 200 CT scans.

2 Vertebrae Localization and Identification

Similar to previous methods, our system relies on a two-stage approach. The first stage aims at roughly detecting and localizing all vertebrae of the spine within the image. Refinement of vertebrae positions and their identification is obtained in the second stage.

Previous work extracts location candidates via classification (*e.g.* in a sliding window framework). In contrast, here we take a more direct, regression approach. In fact, given a training set we learn a regression function which associates vertebrae positions with image points, directly. By combining the predictions of many (possibly sparse) sampled image points, one can obtain robust and efficient

location estimates while avoiding sliding window-like expensive search. Note that these location estimates are not restricted to be inside the visible image domain, and thus, an approximate localization of *all* vertebrae is possible, even if only a small portion of the patient's anatomy is visible.

Contextual reasoning is enabled via long-range spatial features, like the ones used in [14, 15]. This way, the presence of organs such as kidneys, liver, or lungs provide strong indications about the presence of certain vertebrae. Regression forests enable us to select automatically the most discriminative features of accurate prediction. Next, we first formalize the regression technique. Then we describe the second, refinement stage.

2.1 Stage 1: Regression Forests

Regression forests is a supervised learning technique for the probabilistic estimation of continuous variables. Recent work has shown that this technique can be successfully applied to organ bounding box localization in CT [14] and MR [15]. In our application, we aim at regressing the set of n vertebrae centroids denoted as $\mathcal{C}=\{\mathbf{c}_i\}_1^n$ with $\mathbf{c}\in\mathbb{R}^3$. The predictor function is then defined as $p(\mathcal{C}|\mathcal{X})$ where $\mathcal{X}=\{(\mathbf{x}_j,\mathbf{f}_j)\}$ is a set of pairs of feature vectors $\mathbf{f}=(f_1,\ldots,f_d)\in\mathbb{R}^d$ with visual feature responses f extracted for individual image points $\mathbf{x}\in\mathbb{R}^3$. Thus, given the data $\mathcal{X}$ obtained from an image this discriminative predictor allows to estimate the most likely positions of vertebrae in that image.

Regression forests tackle the problem of learning the predictor in a divide-and-conquer fashion. A forest is an ensemble of T (probabilistic) binary decision trees, where each tree t learns its own predictor $p_t(\mathcal{C}|\mathcal{X})$. Given a training set $\mathcal{T}=\{(\mathcal{X}_k,\mathcal{C}_k)\}$, obtained from annotated CT scans, training a tree is done by successively subdividing the training examples within the feature space. At each internal node data subsets $\mathcal{T}_\mathrm{L},\mathcal{T}_\mathrm{R}$ are sent to the left and right child node. A local split function is determined at each node based on the arriving examples. The splits are obtained by (randomly) selecting one feature response for all examples and optimizing over a threshold w.r.t. an objective function. The splitting aims at clustering examples in leaf (terminal) nodes with both consistent annotations and similar feature responses. Tree growing stops when a certain tree depth is reached. In order to extract visual feature responses $\mathbf{f}$, we employ displaced box features which: i) capture long-range appearance context [14, 15] and, ii) can be implemented efficiently via integral image processing [16]. Injecting randomness during the tree training process decreases correlation between individual trees and increases the forest generalization capabilities[1].

Forest Training. Since our training set is composed of feature vectors for image points from arbitrary, unregistered CT scans with varying resolutions and croppings, a regression over absolute image coordinates of vertebrae is not meaningful. Instead, and similar to the case of bounding box regression [14, 15], we associate each training point $\mathbf{x}$ with its relative displacements $\{\mathbf{d}_i\}$, *i.e.* the offsets to *all* available vertebrae centroids given by $\mathbf{d}_i=\mathbf{c}_i-\mathbf{x}$. We employ multivariate

[1] More details on forests can be found in [13].

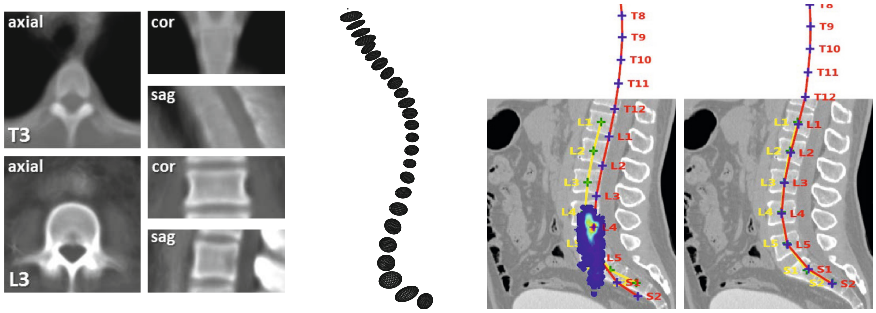

Fig. 1. (a) Mean images of our appearance model for T3 and L3. (b) Visualization of Gaussian densities for offset probabilities in our shape model. Ellipses illustrate one standard deviation w.r.t. covariance matrices. (c) Output of the regression forest for a test image (red) with an overlay of expert annotation (yellow) and prediction distribution for L4. (d) Result after refinement via HMM. Besides accurate predictions for vertebrae within the image, our method yields reasonable predictions outside.

Gaussians to model the node predictor functions: $\mathcal{N}(\mu, \Sigma|\mathcal{D})$ with $\mu \in \mathbb{R}^{3n}$ where $\mathcal{D} = \{\{\mathbf{d}_i\}_j\}$ is the set of offsets obtained from all training points arriving at that node. The training objective function is defined as $\xi(\mathcal{T}_\mathrm{L}, \mathcal{T}_\mathrm{R}) = \mathrm{tr}(\Sigma_\mathrm{L}) + \mathrm{tr}(\Sigma_\mathrm{R})$. Node training aims at minimizing ξ which, in turn, minimizes the diagonal entries of the covariance matrices. This produces child subsets $\mathcal{T}_{\mathrm{L,R}}$ with lower uncertainty and higher confidence in the prediction of vertebrae location.

Forest Testing. Given a previously unseen CT scan, image points cast a (probabilistic) vote on the position of all vertebrae. In fact, each point is pushed through all trained trees. Each split node then sends the point to its left or right child depending on its feature vector, recursively until the point reaches a leaf node. The corresponding predictor function (*i.e.* Gaussian in this case) is read out and used for making one prediction for all vertebrae positions relative to the image point location. Aggregating all predictions over all trees and test points yields a distribution over all vertebrae positions (see Fig.1(c)). For robustness, we approximate the maximum a-posteriori estimate $\hat{\mathcal{C}}$ of this distribution through mean-shift (initialized with the maximum response of a low-resolution histogram over predictions with bin size 4mm). The output vertebrae locations obtained here are then used as input for our refinement step, described next.

2.2 Stage 2: Hidden Markov Model

The second stage of our approach aims at refining the localization of all centroids of vertebrae visible in the image. To this end, we employ a joint prior model of vertebra appearance and spinal shape. The model parameters are optimized using the same data set employed to train the forest.

Vertebra Appearance Model. The appearance model consists of pairs of mean and variance images $\mathcal{A} = \{(M_i, V_i)\}_1^n$, one pair per vertebrae. These pairs

are computed by super-imposing sub-volumes of size $11 \times 11 \times 5$cm cropped from the training data and centered on each annotated vertebra. A few iterations of nonlinear registration are performed to increase the sharpness of the mean images (see examples in Fig. 1(a)). Given a candidate position $\mathbf{c}_i$ we define a likelihood function w.r.t. the appearance model as

$$p(\mathbf{c}_i|\mathcal{A}) = \int_{\Omega_i} \frac{1}{\sqrt{2\pi V_i(\mathbf{x})}} \exp\left(-\frac{(I(\mathbf{c}_i - \mathbf{x}) - M_i(\mathbf{x}))^2}{2V_i(\mathbf{x})}\right) d\mathbf{x} \quad . \tag{1}$$

Spine Shape Prior. The shape model captures conditional probabilities over vertebrae positions. We determine a set of distributions $\mathcal{S} = \{p(\mathbf{c}_i|\mathbf{c}_{i-1}, s)\}_2^n$ where

$$p(\mathbf{c}_i|\mathbf{c}_{i-1}, s) \hat{=} \mathcal{N}\left(\frac{\|\mathbf{c}_i - \mathbf{c}_{i-1}\|}{s}|\mu_i, \Sigma_i\right) \quad \text{with} \quad s = \frac{1}{n-1}\sum_{i=2}^{n} \frac{\|\mathbf{c}_i - \mathbf{c}_{i-1}\|}{\mathbb{E}(\|\mathbf{c}_i - \mathbf{c}_{i-1}\|)} \tag{2}$$

The variable s corresponds to a global scale factor which reflects overall body size. A visualization of these offset distributions is shown in Fig. 1(b).

Joint Shape and Appearance. We define an HMM with hidden states for each vertebrae position, appearance likelihoods and inter-vertebra shape priors. The HMM distribution $p(\mathcal{C}|\mathcal{A}, \mathcal{S}, s)$ conditioned on global scale yields the energy:

$$E(\mathcal{C}; s) = -\sum_{i=1}^{n} \log\left[p(\mathbf{c}_i|\mathcal{A})\right] - \lambda \sum_{i=2}^{n} \log\left[p(\mathbf{c}_i|\mathbf{c}_{i-1}, s)\right] \quad . \tag{3}$$

Given a value for s and multiple sampled location candidates MAP inference can be achieved via dynamic programming. Several thousand candidate locations are sampled from the vicinity of the forest prediction using a normal distribution $\mathcal{N}(\mathbf{c}_i, \sigma^2)$ with $\sigma = 30$mm. In practice, we optimize over 7 scale parameters from a $[0.85, 1.15]$ interval covering 97% of observed patient scales. The weighting parameter λ controls the influence of the shape term, and thus, how much the solution can deviate from the mean shape. Throughout our experiments this weighting is fixed to 0.1. In the exemplary result in Fig. 1(d) notice how the thoracic vertebrae follow reasonable predictions outside the image domain.

3 Experiments

Our spine model includes $n = 26$ individual vertebrae, where the regular 24 from the cervical, thoracic, and lumbar regions are augmented with 2 centroids denoted as S1 and S2 located on the sacrum. We evaluate accuracy of both localization and identification on a dataset of 200 CT scans where the centroids of all visible vertebrae have been manually selected. The dataset is a heterogeneous collection of CT scans from different clinical centers equipped with varying hardware. Images have been acquired for diverse clinical tasks. The scans vary widely, especially in terms of vertical cropping, image noise and physical resolution. The inter-axial distance varies between 0.5 and 6.5mm, with 79 scans having a distance of 3.75mm. The number of slices varies between 51 and 2058 with an average of about 240. Some highly cropped images show only 4 vertebrae.

Table 1. Summary of the localization and identification errors evaluated on 200 CTs

Vertebrae		Stage 1: Regression Forest			Stage 2: HMM			Distance to Closest			Identification	
Region	Counts	Median	Mean	Std	Median	Mean	Std	Median	Mean	Std	Correct	Rate
All	2595	15.91	18.35	11.32	5.31	9.50	10.55	4.79	6.10	5.53	2089	81%
Cervical	116	25.97	30.74	18.64	6.87	10.85	12.49	6.14	8.53	9.05	84	72%
Thoracic	1417	15.79	18.20	10.81	5.51	9.83	10.44	4.91	5.94	4.84	1100	78%
Lumbar	1062	15.40	17.20	10.07	4.88	8.92	10.45	4.59	6.06	5.82	905	85%

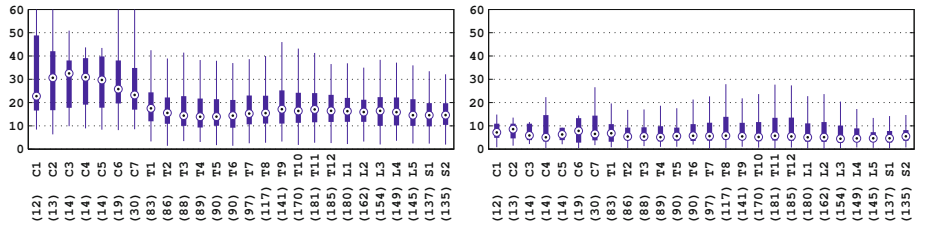

Fig. 2. Error statistics for individual vertebrae: (left) forest prediction only, (right) refinement via HMM. The counts of each vertebra in our database are given in brackets.

3.1 Results

We split the 200 CT scans into two non-overlapping sets with 100 scans each. Each set is used once for both: i) forest training (50 trees, depth 20), and ii) estimating the shape and appearance model; the remaining set is used for testing. Thus we can report errors for all 200 scans and a total of 2595 vertebrae.

Localization Errors defined as distance (in mm) of each predicted vertebra location from its expert annotation are summarized in Tab. 1 and Fig. 2. We obtain a median error of less than 6mm. The highest errors are within the cervical region with a median of about 7mm. This is due to the low number of cervical vertebrae in our data sets (only 6-17 examples for C1 to C7). The lowest errors are obtained for the lumbar region (including S1 and S2) where visual appearance is more discriminative and low image resolution has less of an impact. In Fig. 2, we plot the statistics over localization errors graphically. The figure highlights the massive improvement due to the HMM refinement step. We also give the counts for vertebrae as they appear in the set of 200 scans.

In Tab. 1 we also report distances of predictions from the closest vertebra. So, we have an estimate whether our prediction is in fact located on a vertebra, even if it is not the correct one. The difference between these errors and the ones when considering the correct vertebra is higher for the cervical region, while in the thoracic and lumbar region the difference between median errors is less than 0.6mm. This indicates that in most cases the closest centroid in T and L regions is the correct one, while in C our predicted localizations are in fact on the spine, but might be in some cases on the incorrect vertebra. This confirms that the increased difficulty in discriminating close-by vertebrae in the cervical and upper

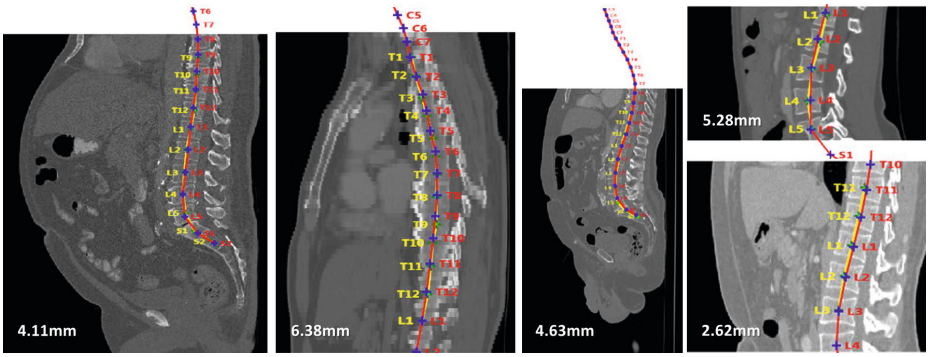

Fig. 3. Our results (red) for varying CT scans (cropped, low resolution, noise, and large field-of-view). Numbers are the mean errors w.r.t. expert annotations (yellow).

thoracic regions contributes to most of our errors. Still, even in these challenging cases our system is able to robustly localize the overall spinal anatomy, which in certain applications might be sufficient.

Identification Errors. We define a vertebra identification criterion as follows: if the closest centroid in the expert annotation corresponds to the correct vertebra, and the localization error is less than 20mm, we call the identification correct. The last two columns in Tab. 1 show an overall success rate of 81%, *i.e.* 2089 out of 2595 vertebrae correctly identified.

Efficiency. In the proposed system training a single tree takes about 3 minutes on randomly sampled 5% of the image points of 100 scans. Each tree can be trained independently and in parallel. More importantly, testing is very fast. In fact, testing a whole forest on a scan takes less than 1 second. The HMM-based refinement takes about 5-15 seconds for each scale s, also depending on the number of vertebrae within the image. Thus, in total, the localization and identification of all vertebrae in one test image is achieved in less than 2 minutes. Figure 3 provides some visual results and corresponding mean errors.

4 Conclusion

This paper has proposed an automatic and efficient approach for the localization and identification of vertebrae in generic CT scans. The algorithm does not make any assumptions on the input images and can deal with highly cropped scans and partially visible spines. Exhaustive experiments on a database of 200 labelled CT scans demonstrate the strength of our joint model of discriminative regression and generative appearance and shape modeling.

In the future, increasing the amount of training data, in particular, for the cervical region would produce an increase in accuracy across the entire spine.

Additionally, automatically predicting the patient overall size could replace the current scale search step and reduce testing times to only a few seconds. Further investigation will also be carried out w.r.t. highly pathological cases of spine such as high-grade scoliosis and cifosis.

References

1. Peng, Z., Zhong, J., Wee, W., Lee, J.H.: Automated Vertebra Detection and Segmentation from the Whole Spine MR Images. In: IEEE EMBC, pp. 2527–2530 (2005)
2. Pekar, V., Bystrov, D., Heese, H.S., Dries, S.P.M., Schmidt, S., Grewer, R., den Harder, C.J., Bergmans, R.C., Simonetti, A.W., van Muiswinkel, A.M.: Automated Planning of Scan Geometries in Spine MRI Scans. In: Ayache, N., Ourselin, S., Maeder, A. (eds.) MICCAI 2007, Part I. LNCS, vol. 4791, pp. 601–608. Springer, Heidelberg (2007)
3. Schmidt, S., Kappes, J.H., Bergtholdt, M., Pekar, V., Dries, S.P.M., Bystrov, D., Schnörr, C.: Spine Detection and Labeling Using a Parts-Based Graphical Model. In: Karssemeijer, N., Lelieveldt, B. (eds.) IPMI 2007. LNCS, vol. 4584, pp. 122–133. Springer, Heidelberg (2007)
4. Huang, S.H., Lai, S.H., Carol, L.N.: A statistical learning approach to vertebra detection and segmentation from spinal MRI. In: IEEE ISBI, vol. 28, pp. 1595–1605 (2008)
5. Huang, S.H., Chu, Y.H., Lai, S.H., Novak, C.L.: Learning-based Vertebra Detection and Iterative Normalized-Cut Segmentation for Spinal MRI. IEEE TMI 28(10), 1595–1605 (2009)
6. Kelm, B., Zhou, K., Sühling, M., Zheng, Y., Wels, M., Comaniciu, D.: Detection of 3D Spinal Geometry Using Iterated Marginal Space Learning. In: Workshop MedCV (2010)
7. Ma, J., Lu, L., Zhan, Y., Zhou, X., Salganicoff, M., Krishnan, A.: Hierarchical Segmentation and Identification of Thoracic Vertebra Using Learning-Based Edge Detection and Coarse-to-Fine Deformable Model. In: Jiang, T., Navab, N., Pluim, J.P.W., Viergever, M.A. (eds.) MICCAI 2010, Part I. LNCS, vol. 6361, pp. 19–27. Springer, Heidelberg (2010)
8. Oktay, A.B., Akgul, Y.S.: Localization of the Lumbar Discs Using Machine Learning and Exact Probabilistic Inference. In: Fichtinger, G., Martel, A., Peters, T. (eds.) MICCAI 2011, Part III. LNCS, vol. 6893, pp. 158–165. Springer, Heidelberg (2011)
9. Klinder, T., Ostermann, J., Ehm, M., Franz, A., Kneser, R., Lorenz, C.: Automated Model-based Vertebra Detection, Identification, and Segmentation in CT Images. MedIA 13(3), 471–482 (2009)
10. Bishop, C.: Pattern Recognition and Machine Learning. Springer (2006)
11. Corso, J.J., Alomari, R.S., Chaudhary, V.: Lumbar Disc Localization and Labeling with a Probabilistic Model on Both Pixel and Object Features. In: Metaxas, D., Axel, L., Fichtinger, G., Székely, G. (eds.) MICCAI 2008, Part I. LNCS, vol. 5241, pp. 202–210. Springer, Heidelberg (2008)
12. Alomari, R., Corso, J., Chaudhary, V.: Labeling of Lumbar Discs using both Pixel- and Object-level Features with a Two-Level Probabilistic Model. IEEE TMI 30(1), 1–10 (2011)

13. Criminisi, A., Shotton, J., Konukoglu, E.: Decision Forests: A Unified Framework. Foundations and Trends in Computer Graphics and Vision 7(2-3) (2011)
14. Criminisi, A., Shotton, J., Robertson, D., Konukoglu, E.: Regression Forests for Efficient Anatomy Detection and Localization in CT Studies. In: Workshop MedCV (2010)
15. Pauly, O., Glocker, B., Criminisi, A., Mateus, D., Möller, A.M., Nekolla, S., Navab, N.: Fast Multiple Organ Detection and Localization in Whole-Body MR Dixon Sequences. In: Fichtinger, G., Martel, A., Peters, T. (eds.) MICCAI 2011, Part III. LNCS, vol. 6893, pp. 239–247. Springer, Heidelberg (2011)
16. Viola, P., Jones, M.J.: Robust Real-Time Face Detection. IJCV 57(2), 137–154 (2004)