Automatic Peripheral Nerve Segmentation in Presence of Multiple Annotators

Julián Gil González(✉), Andrés M. Álvarez, Andrés F. Valencia, and Álvaro A. Orozco

Faculty of Engineering, Universidad Tecnológica de Pereira, Pereira 660003, Colombia
{jugil,andres.alvarez1,andres.valencia,aaog}@utp.edu.co

Abstract. Peripheral Nerve Blocking (PNB) is a technique commonly used to perform regional anesthesia. The success of PNB procedures lies of the accurate location of the target nerve. The ultrasound images (UI) have frequently been used aiming to locate nerve structures in the context of PNB procedures. This type of images allows a direct visualization of the target nerve, and the anatomical structures around it. Notwithstanding, the nerve segmentation in UI by an anesthesiologist is not straightforward since these images are affected by several artifacts; hence, the accuracy of nerve segmentation depends on the anesthesiologist expertise. In this sense, we face a scenario where we have manual multiple nerve segmentations performed by several anesthesiologists with different levels of expertise. In this paper, we propose a nerve segmentation approach based on supervised learning. For the classification step, we compare two schemes based on the concepts "Learning from crowds" aiming to code the information of multiple manual segmentations. Attained results show that our approach finds a suitable UI approximation by ensuring the identification of discriminative nerve patterns according to the opinions given by multiple specialists.

1 Introduction

Recently, regional anesthesia has become an attractive alternative for general anesthesia in the context of medical surgeries, mainly because it improves postoperative mobility and reduces morbidity and mortality [1]. Regional anesthesia comprises the administration of an anesthetic substance in the area surrounding a nerve structure to block the transmission of nociceptive information (this procedure is known as Peripheral Nerve Blocking, PNB) [2]. In this regard, the success of regional anesthesia depends on the accurate location of the target nerve [3]. The use of ultrasound images (UI) has gained considerable interest to locate nerve structures in PNB procedures [1]. This method allows a direct visualization of the target nerve and the anatomical structures around it [4]. Notwithstanding, the localization of nerve structures in ultrasound images is a challenging task for the specialists (in this case anesthesiologists) due to these kind of images are affected by several artifacts such as attenuation, acoustic

M. Mendoza and S. Velastín (Eds.): CIARP 2017, LNCS 10657, pp. 246–254, 2018.
https://doi.org/10.1007/978-3-319-75193-1_30

shadows, and speckle noise [5]. Thus, the accurate delimitation of a given target depends on the operator (anesthesiologist) experience [6].

The above problem can be minimized using automatic nerve-segmentation systems, which are intended to assist the anesthesiologist with the aim of locating nerve structures in PNB procedures. Nevertheless, to build a system with these specifications, it is necessary to access to the actual label, that is, we need ultrasound images indicating which regions correspond to a nerve (usually, this process is performed by an anesthesiologist) [1,6]. In practice, the above is considered as a problem since it is not possible for the specialists to accurately identify nerve structures in an ultrasound image, considering that the speckle noise and the artifacts difficult the delimitation of anatomical structures [7]. In this sense, the obtained labels do not correspond to the Ground Truth but a subjective interpretation (possibly noisy) given by the specialist based on his experience and his training. For the automatic segmentation of nerve structures, without the knowledge of the Ground Truth, a set of noisy annotations (manual segmentation) from various specialists could be used. In this case, it is necessary to use the manual segmentation provided by multiple experts with the aim of building a segmentation model that allows to measure the performance of the annotators based on the parametrization of the ultrasound images to deal with the subjectivity presented in the labeled regions.

In the presence of multiple annotators, the labels from several experts have been used in different ways with the aim of building automatic systems for nerve segmentation. For example, in [6], the authors consider as the gold standard the annotations from one specialist. On the other hand, in [1] the authors use the Majority voting from the annotations as the ground truth. However, these approaches have some problems, for instance, if we only use the labels from one of the annotators, the segmentations results would be biased by the expertise of the annotators. Similarly, in the Majority Voting approach, it is considered that all the annotators are equally reliable, which is not common in real scenarios [8]. Another way to deal with the problem of not having the gold standard is to use a recent trend in machine learning named “Learning from multiple annotators”. The area of learning with multiple annotators is relatively new; its aim is to perform supervised learning task when the gold standard is not available, and we just have access to multiple annotations provided by several experts or annotators. This area has been applied to problems such as, regression [9], classification [10], and sequence labeling [11].

In this paper, we present a method for the automatic segmentation of nerve structures depicted in ultrasound images considering the scenario where the ground truth is not available. In particular, we use two classification schemes with multiple annotators aiming to combine the manual segmentation from different experts to reveal discriminant patterns associated with the nerve structures. One of them is based on Logistic Regression, where the annotator performance depends only on the true label and is measured in terms of sensitivity and specificity [12]. The second scheme is a model based on Logistic Regression, where it is assumed that the annotator performance depends on both, the true label and the instance that the annotator is labeling [13]. We hypothesize that by using classifications schemes considering that consider the non-availability of the ground truth, it is possible to reduce the subjectivity present in the labeled regions. There are a few works that consider the information of several annotators to build nerve-segmentation systems [1,6]. However, these works use basic schemes to deal with this information (majority voting or an approach where only the information from one annotator is considered), these approaches are not suitable since they consider that the experts have the same level of expertise. In this sense, the main contribution of our work lies in to develop an automatic nerve-segmentation system, which captures the segmentation expertise from different specialist considering a non-homogeneity in the performance of experts. The obtained results show that our approach finds a suitable UI approximation by ensuring the identification of discriminative nerve patterns according to the opinions given by multiple specialists. Indeed, our proposal outperforms state-of-the-art approaches that carry out nerve segmentation in terms of a Dice coefficient assessment.

2 Materials and Methods

2.1 Multi-annotator Classifications Schemes

For the training of a typical classification problem (i.e. a classification scheme without considering multiple annotators) we dispose a training set $\mathscr{D} = \{(\mathbf{x}_i, t_i)\}_{i=1}^{N}$, with N samples, where $\mathbf{x}_i$ is an instance known as the $D-$dimensional feature vector and t_i is the label associated to x_i, which is assumed as the "ground truth". However, in this work, we take into account the case where the ground-truth is not available for the training, and in contrast, we only have access to an amount of labels (possibly noisy) provided by R experts o annotators [12]. In this regard, the training set in the context of multiple annotators is $\mathscr{D} = \{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^{N}$, where $\mathbf{y}_i = y_i^1, \ldots, y_i^R$ are the annotations for the i-th sample given by the R annotators. In this work, we use two classification schemes with multiple annotators to deal with the problem of automatic segmentation of nerve structures. The following is a brief description of these methods, where they establish a random variable $\mathbf{z} = [z_1, \ldots, z_N]$, which represents the unknown ground-truth for the i-th sample.

Logistic regression with multiple annotators (LFC). We follow the multi-annotator classification model proposed in [12]. The annotator performance is measured in terms of sensitivity α^r and specificity β^r, where $\alpha^r = p\left(y^r = 1 | z = 1\right),\ \beta^r = p\left(y^r = 0 | z = 0\right).$ Hence, we use the training dataset to construct a multiple-annotator classification based on logistic regression [14]. In this sense, given the samples and the annotations, we need to estimate the parameters associated with the performance of each annotator $\boldsymbol{\alpha} = \left[\alpha^1, \ldots, \alpha^R\right]$,

$\boldsymbol{\beta} = \left[\beta^1, \ldots, \beta^R\right]$, and the parameters associated with the classifier $\boldsymbol{w}$. For estimating these parameters, we employ an Expectation-Maximization (EM) algorithm. The likelihood function is given as

$$p\left(\mathscr{D}|\boldsymbol{\theta}\right) = \prod_{i=1}^{N} \left[p_i \prod_{r=1}^{R} \left(\alpha^r\right)^{\left(y_i^r\right)} \left(1-\alpha^r\right)^{\left(1-y_i^r\right)} + \left(1-p_i\right) \prod_{r=1}^{R} \left(\beta^r\right)^{\left(1-y_i^r\right)} \left(1-\beta^r\right)^{\left(y_i^r\right)} \right],$$

where $\boldsymbol{\theta} = \{\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{w}\}$, and p_i is computed by means of a "Logistic Regression" function [14]. The EM algorithm is performed from the following steps:

E-step: The conditional expectation of the log-likelihood yields

$$\mathbb{E}[\ln(p(\mathscr{D}, \mathbf{z}|\boldsymbol{\theta}))] = \sum_{i=1}^{N} \mathbb{E}[z_i] \ln(a_i p_i) + (1 - \mathbb{E}[z_i]) \ln\left(b_i(1-p_i)\right),$$

where $a_i = \prod_{r=1}^{R} \left(\alpha^r\right)^{\left(y_i^r\right)} \left(1-\alpha^r\right)^{\left(1-y_i^r\right)}$, $b_i = \prod_{r=1}^{R} \left(\beta^r\right)^{\left(1-y_i^r\right)} \left(1-\beta^r\right)^{\left(y_i^r\right)}$, and $\mathbb{E}[z_i]$ is the estimated ground truth which follows $\mathbb{E}[z_i] = \mu_i = \dfrac{a_i p_i}{a_i p_i + b_i(1-p_i)}$.

M-step: Given the estimated gold standard μ_i and the training data, we estimate the parameters $\boldsymbol{\theta}$ by maximizing the conditional expectation of the log-likelihood computed in the E-step. The annotators performance parameters are updated using

$$\alpha^r = \frac{\sum_{i=1}^{N} \mu_i y_i^r}{\sum_{i=1}^{N} \mu_i}, \quad \beta^r = \frac{\sum_{i=1}^{N} \left(1-\mu_i\right)\left(1-y_i^r\right)}{\sum_{i=1}^{N} \left(1-\mu_i\right)}.$$

Finally, the parameters related with the logistic regression classifier, can be calculated by using similar equations to the single annotator context, where the true labels are changed for soft labels given by μ_i. See [14].

Modeling annotator expertise: Learning when everybody knows about something (MAE). We follow the multi-annotator classification schemes proposed in [13]. This model is an extension of the proposed model in [12]. Unlike the model **LFC**, the model **MAE** consider that the label given by the annotator r depends on the unknown true label z_i and the instance $\mathbf{x}_i$ that he is labeling, in this sense

$$p\left(y_i^r|\mathbf{x}_i, z_i\right) = \left(1-\eta_r(\mathbf{x}_i)\right)^{|y_i^r - z_i|} \eta_r(\mathbf{x}_i)^{1-|y_i^r - z_i|},$$

where $\eta_r(\mathbf{x}_i)$ follows a Logistic regression model $\eta_r(\mathbf{x}_i) = \left(1 + \exp\left(-\boldsymbol{\lambda}_r^\top \mathbf{x}_i\right)\right)^{-1}$. Given the dataset, we need to estimate the parameters associated with the performance of each annotator $\boldsymbol{\Lambda} = [\boldsymbol{\lambda}_1, \ldots, \boldsymbol{\lambda}_R]$ and the parameters associated with the classifier $\boldsymbol{w}$ based on "Logistic Regression" [14]. For estimating these parameters, we employ an Expectation-Maximization (EM) algorithm. The likelihood function is given as

$$p\left(\mathscr{D}|\boldsymbol{\phi}\right) = \prod_{i=1}^{N} \left[p_i \prod_{r=1}^{R} \left(1-\eta_r(\mathbf{x}_i)\right)^{1-y_i^r} \eta_r(\mathbf{x}_i)^{y_i^r} + \left(1-p_i\right) \prod_{r=1}^{R} \left(1-\eta_r(\mathbf{x}_i)\right)^{y_i^r} \eta_r(\mathbf{x}_i)^{1-y_i^r} \right],$$

where $\boldsymbol{\phi} = \{\boldsymbol{\Lambda}, \boldsymbol{w}\}$, and p_i is computed by means of a "Logistic Regression" function [14]. The EM algorithm is performed from the following steps:

E-step: The conditional expectation of the log-likelihood is defined as

$$\mathbb{E}[\ln(p(\mathscr{D}, \mathbf{z}|\boldsymbol{\phi}))] = \sum_{i=1}^{N} \mathbb{E}[z_i] \ln(c_i p_i) + (1 - \mathbb{E}[z_i]) \ln\left(d_i(1 - p_i)\right),$$

where $c_i = \prod_{r=1}^{R} (1 - \eta_r(\mathbf{x}_i))^{1-y_i^r} \, \eta_r(\mathbf{x}_i)^{y_i^r}$, and $d_i = \prod_{r=1}^{R} (1 - \eta_r(\mathbf{x}_i))^{y_i^r} \, \eta_r(\mathbf{x}_i)^{1-y_i^r}$, and $\mathbb{E}[z_i]$ is the estimated ground truth which follows $\mathbb{E}[z_i] = \mu_i = \dfrac{c_i p_i}{c_i p_i + d_i(1 - p_i)}$.

M-step: Given the estimated gold standard μ_i and the training data, we estimate the parameters $\boldsymbol{\phi}$ by maximizing The conditional expectation of the log-likelihood computed in the E-step. To compute the parameters $\boldsymbol{\lambda}$ related to the model, we use gradient-based methods. Next we provided the first order derivate w.r.t. $\boldsymbol{\lambda}$

$$\frac{\partial \mathbb{E}[\ln(p(\mathscr{D}, \mathbf{Z}|\boldsymbol{\theta}))]}{\partial \boldsymbol{\lambda}_r} = \sum_{i=1}^{N} (-1)^{y_i^r} \left(1 - 2\mu_i\right) \eta_r(\mathbf{x}_i) \left(1 - \eta_r(\mathbf{x}_i)\right) \mathbf{x}_i$$

Finally, the parameters related with the logistic regression classifier can be calculated by using similar equations to the single annotator context, where the true labels are changed for soft labels given by μ_i. See [14].

3 Results and Discussions

Ultrasound imaging dataset: To validate the nerve segmentation approach based on classification with multiple annotators, we use a dataset named *UI-UTP*, which consists of recordings of ultrasound images from patients who underwent regional anesthesia using the Peripheral nerve blocking procedure. This dataset is composed of 48 ultrasound images from the ulnar nerve (21 images) and median nerve (27 images). Each ultrasound image was collected using a Sonosite Nano-Maxx device (the resolution of each image is 640×480 pixels). Each image in the dataset was labeled by three specialists in anesthesiology to indicate the location of the nerve structures.

Segmentation Scheme training and testing: A leave-one-out validation scheme is employed to compute the system performance regarding the nerve segmentation in the context of multiple annotators. The nerve segmentation considering multiple experts comprises the following stages: First, we use Graph Cuts Segmentation [15] to define a region of interest (ROI) in which the nerve region is probably located. Then, a median filter is applied over the ROIs to reduce the speckle noise effect while enhancing the UI quality. Then, each filtered image is divided into different regions by using SLIC-superpixel [16], and each of these superpixel is parametrized from the non-linear Wavelet transform (for details, see [6]). Now, in this work, the segmentation problem is considered as a binary classification, where each parametrized superpixel is classified as nerve or background. However, as we have previously pointed out, it is not possible to obtain the ground truth (i.e. the labels indicating which superpixel is a nerve region and what it is not) since in this case, the labels correspond to a subjective assessment given by an anesthesiologist. According to the above, we use two schemes for binary classification in the context of multiple annotators, one of them is a model based on Logistic Regression, where the annotator performance depends on the true label and is measured in terms of sensitivity and specificity [12] (**LFC**). The second scheme is a model based on Logistic Regression, where it is assumed that the annotator performance depends on both, the true label and the instance that the annotator is labeling [13] (**MAE**). Finally, we use a methodology based on morphological operators aiming to improve the segmentations results (see [6]). The system performance is measured in terms of the Dice coefficient (DC) that quantifies the overlap between the UI segmented based on the proposed segmentation approach and the label given by the specialists. In addition, we consider two common approaches to deal classification problems with multiple annotators, the first approach is to use and scheme based on Logistic Regression, where the labels from each one of the annotators are considered as gold standard (**LR-EX1**, **LR-EX2**, **LR-EX3**); the second methodology is based on Logistic Regression, where we consider as true labels the majority voting from the annotations (**LR-MV**).

Obtained results: To visually compare the attained nerve identification in the context of multiple experts, Fig. 1 exposes some segmentation results regarding each multi-annotator classification approach. Overall, all classification methods allow to identify relevant patterns to segment nerves from the UI. Nevertheless, due to the absence of the true label, typical classifications methods (**LR-EX1**, **LR-EX2**, **LR-EX3**, and **LR-MV**) fail in the complete identification of nerves by generating false negatives (i.e., classify as background nerve regions). The above is a significant issue since the anesthesiologist needs an accurate delimitation of the nerve structure aiming to define the point where the anesthetic should be spread. Unlike these classification methods, our approach based on multiple annotators (specifically **MAE**) reduces the number of false negatives considerably in the segmented image offering a better identification of the nerve structure. Table 1 shows the results of the morphological validation in terms of the DC for the leave-one-out validation scheme. We perform a statistical significance analysis based on the equal mean test for the segmentation approaches considered in this work. This test allows to determine which method provides a higher performance in terms of the Dice coefficient. From the results exposed, the segmentation scheme based on multiple annotators (specifically the multiple annotators model proposed in [13]) outperforms state-of-the-art approaches which are based on typical supervised learning schemes (i.e. supervised learning approaches without considering multiple annotators). These results can be explained in the sense that the multiple annotators schemes are based on the

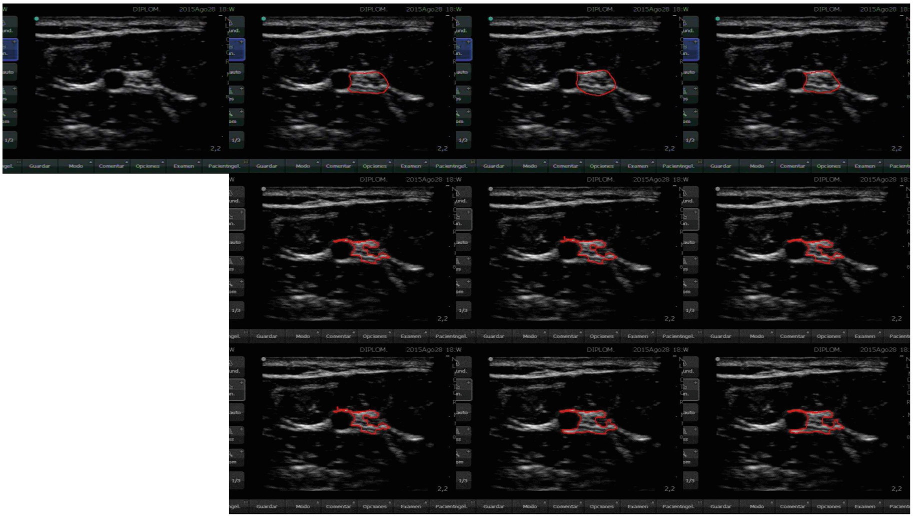

Fig. 1. Segmentation results for an ulnar nerve. On the top from left to right, we show the original image and the labels provided by the three experts. On the second row from left to right, we expose the segmentation results provided by **LR-EX1**, **LR-EX2**, and **LR-EX3**. Finally, on the bottom from left to right we show the segmentation results for **LR-MV**, **LFC**, and **MAE**.

Table 1. Nerve segmentation validation in terms of the Dice coefficient.

	Dice coefficient $\mu \pm \sigma$
LR-Ex1	0.6380 ± 0.0035
LR-Ex2	0.6536 ± 0.0043
LR-Ex3	0.6304 ± 0.0042
LR-MV	0.6414 ± 0.0052
LFC [12]	0.6446 ± 0.0011
MAE [13]	**0.6557 ± 0.0015**

fact that the gold standard is not available for the training stage, where the true label is estimated from the manual segmentations provided by different experts in anesthesiology. In contrast, segmentation approaches based on typical supervised learning assume as a gold standard the manual segmentations from one of the annotators, which implies that the classifier predictions will be biased according to annotator expertise.

4 Conclusion

In this paper, we discuss a first attempt for the design of nerve-segmentation systems based on classification with multiple annotators. In this sense, we perform the nerve identification by considering the case where the ground truth is not available. In fact, this consideration is not far from reality since the nerve identification in UI depends on the specialist expertise. So, we use multi-annotator classification schemes to estimate the unavailable true label and the classifier parameters jointly. We tested our strategy in a real-world nerve segmentation dataset captured by "The Automatics Research Group-Universidad Tecnológica de Pereira," which holds UI images of ulnar and median nerves. The experimental results showed that the segmentation methodology based on the information from different experts outperforms state-of-the-art alternatives for nerve segmentation in terms of the Dice coefficients. Hence, the proposed method have a better interpretation of the patterns associated with the nerves by combining the manual segmentation given by multiple anesthesiologists. As future work, authors plan to use more robust multi-annotators classification schemes (for example approaches based on deep-learning) to improve further the quality of the nerve segmentation.

Acknowledgment. This work was funded by Colciencias under the project: "Desarrollo de un sistema de identificación de estructuras nerviosas en imágenes de ultrasonido para la asistencia de bloqueo de nervios periféricos. Aplicación al tratamiento de dolor agudo traumático y prevención del dolor neuropático crónico" (code: 1110-744-55958).

References

1. Hadjerci, O., Hafiane, A., Conte, D., Makris, P., Vieyres, P., Delbos, A.: Computer-aided detection system for nerve identification using ultrasound images: a comparative study. Inform. Med. Unlocked **3**, 29–43 (2016)
2. Hadjerci, O., Hafiane, A., Makris, P., Conte, D., Vieyres, P., Delbos, A.: Nerve detection in ultrasound images using median Gabor binary pattern. In: Campilho, A., Kamel, M. (eds.) ICIAR 2014. LNCS, vol. 8815, pp. 132–140. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-11755-3_15
3. Denny, N.M., Harrop-Griffiths, W.: Editorial I: location, location, location! ultrasound imaging in regional anaesthesia. Br. J. Anaesth. **94**(1), 1–3 (2005)
4. Shi, J., Schwaiger, J., Lueth, T.C.: Nerve block using a navigation system and ultrasound imaging for regional anesthesia. In: 2011 Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBC, pp. 1153–1156. IEEE (2011)
5. Noble, J.A., Boukerroui, D.: Ultrasound image segmentation: a survey. IEEE Trans. Med. Imaging **25**(8), 987–1010 (2006)
6. González, J.G., Álvarez, M.A., Orozco, Á.A.: A probabilistic framework based on SLIC-superpixel and Gaussian processes for segmenting nerves in ultrasound imagess. In: 2016 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), pp. 4133–4136. IEEE (2016)
7. Chan, V.W.S., Nova, H., Abbas, S., McCartney, C.J.L., Perlas, A., et al.: Ultrasound examination and localization of the sciatic nerve: a volunteer study. J. Am. Soc. Anesth. **104**(2), 309–314 (2006)
8. Rodrigues, F., Pereira, F., Ribeiro, B.: Learning from multiple annotators: distinguishing good from random labelers. Pattern Recogn. Lett. **34**(12), 1428–1436 (2013)

9. Groot, P., Birlutiu, A., Heskes, T.: Learning from multiple annotators with Gaussian processes. In: Honkela, T., Duch, W., Girolami, M., Kaski, S. (eds.) ICANN 2011. LNCS, vol. 6792, pp. 159–164. Springer, Heidelberg (2011). https://doi.org/10.1007/978-3-642-21738-8_21
10. Rodrigues, F., Pereira, F.C., Ribeiro, B.: Gaussian process classification and active learning with multiple annotators. In: ICML, pp. 433–441 (2014)
11. Rodrigues, F., Pereira, F., Ribeiro, B.: Sequence labeling with multiple annotators. Mach. Learn. **95**(2), 165–181 (2014)
12. Raykar, V.C., Yu, S., Zhao, L.H., Valadez, G.H., Florin, C., Bogoni, L., Moy, L.: Learning from crowds. J. Mach. Learn. Res. **11**(Apr), 1297–1322 (2010)
13. Yan, Y., Rosales, R., Fung, G., Schmidt, M.W., Valadez, G.H., Bogoni, L., Moy, L., Dy, J.G.: Modeling annotator expertise: learning when everybody knows a bit of something. In: AISTATS, pp. 932–939 (2010)
14. Bishop, C.M.: Pattern Recogn. Mach. Learn. **128**, 1–58 (2006)
15. Boykov, Y., Funka-Lea, G.: Graph cuts and efficient ND image segmentation. Int. J. Comput. Vis. **70**(2), 109–131 (2006)
16. Achanta, R., Shaji, A., Smith, K., Lucchi, A., Fua, P., Susstrunk, S.: SLIC superpixels. Department and School of Computer and Communication Sciences, EPFL, Lausanne, Switzerland, Technical report, 149300 (2010)