Bioinformatics, 31(18), 2015, 2981–2988
doi: 10.1093/bioinformatics/btv318
Advance Access Publication Date: 20 May 2015
Original Paper

OXFORD

Structural bioinformatics

Computer vision-based automated peak picking applied to protein NMR spectra

Piotr Klukowski[1,†], Michal J. Walczak[2,*,†], Adam Gonczarek[1,*,†], Julien Boudet[2] and Gerhard Wider[2,*]

[1]Department of Computer Science, Wroclaw University of Technology, Wroclaw, Poland and [2]Institute of Molecular Biology and Biophysics, ETH Zurich, 8093 Zurich, Switzerland

Associate Editor: Anna Tramontano

[†]The authors wish it to be known that, in their opinion, the first three authors should be regarded as Joint First Authors.
*To whom correspondence should be addressed.

Received on December 28, 2014; revised on April 27, 2015; accepted on May 18, 2015

Abstract

Motivation: A detailed analysis of multidimensional NMR spectra of macromolecules requires the identification of individual resonances (peaks). This task can be tedious and time-consuming and often requires support by experienced users. Automated peak picking algorithms were introduced more than 25 years ago, but there are still major deficiencies/flaws that often prevent complete and error free peak picking of biological macromolecule spectra. The major challenges of automated peak picking algorithms is both the distinction of artifacts from real peaks particularly from those with irregular shapes and also picking peaks in spectral regions with overlapping resonances which are very hard to resolve by existing computer algorithms. In both of these cases a visual inspection approach could be more effective than a 'blind' algorithm.

Results: We present a novel approach using computer vision (CV) methodology which could be better adapted to the problem of peak recognition. After suitable 'training' we successfully applied the CV algorithm to spectra of medium-sized soluble proteins up to molecular weights of 26 kDa and to a 130 kDa complex of a tetrameric membrane protein in detergent micelles. Our CV approach outperforms commonly used programs. With suitable training datasets the application of the presented method can be extended to automated peak picking in multidimensional spectra of nucleic acids or carbohydrates and adapted to solid-state NMR spectra.

Availability and implementation: CV-Peak Picker is available upon request from the authors.

Contact: gsw@mol.biol.ethz.ch; michal.walczak@mol.biol.ethz.ch; adam.gonczarek@pwr.edu.pl

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Nuclear magnetic resonance (NMR) spectroscopy has become a standard technique in biological research. NMR applied on solutions of biological macromolecules provides data on their structural as well as dynamic properties and supplies detailed information on molecular interactions. The analysis of the NMR spectra often still includes substantial manual work even though a broad range of automated procedures have been developed (Antz *et al.*, 1995; Carrara *et al.*, 1993; Cieslar *et al.*, 1988; Garrett *et al.*, 1991; Güntert, 2004; Herrmann *et al.*, 2002; Herrmann *et al.*, 2002; Hiller *et al.*, 2008; Jung and Zweckstetter, 2004; Kleywegt *et al.*, 1990). Automation of spectral analysis is particularly important in NMR-based drug discovery where hundreds of two-dimensional spectra are measured during screening of libraries of chemical compounds (Coles *et al.*, 2003; Hajduk *et al.*, 1999; Pellecchia *et al.*, 2004; Pellecchia *et al.*, 2008). Moreover, the increasing amount of structural genomics and proteomics using NMR necessitate fast automated procedures to alleviate the time-consuming

user-dependent classical analysis of spectra (Banci *et al.*, 2010; Baran *et al.*, 2004; Parsons and Orban, 2004; Yee *et al.*, 2006). A rapid and robust user-independent pipeline in NMR data analysis would relieve many NMR spectroscopists from routine tasks.

The first step in a fully computerized analysis of NMR spectra is automated peak picking - a task that has attracted substantial interest and there are numerous algorithms available (Abbas *et al.*, 2013; Alipanahi *et al.*, 2009; Antz *et al.*, 1995; Carrara *et al.*, 1993;Cheng *et al.*, 2014; Garrett *et al.*, 1991; Herrmann *et al.*, 2002; Hiller *et al.*, 2008; Kleywegt *et al.*, 1990; Koradi *et al.*, 1998; Liu *et al.*, 2012; Tikole *et al.*, 2014). And yet manual peak picking usually still provides superior results. Existing programs for automated peak picking can suffer from limited ability to discriminate between signals, artifacts and noise requiring external intervention by experienced spectroscopists for an exhaustive recognition of all resonances.

In manual peak picking a researcher visually selects peaks with the expected shapes that, from experience, represent real signals. Thus, he relies on his prior knowledge about the appearance of real peaks as local extrema in the spectrum. An analogous approach is used by people to recognize solid objects in planar images, where small patches are analyzed for the decision whether or not they contain an object. Computer vision is a rapidly developing branch of computer science that attempts to automate this process known as the object detection problem. Computer vision techniques have been successfully applied in many areas as e.g. in face detection (Ahonen *et al.*, 2006; Viola and Jones, 2004), for pedestrian detection (Sabzmeydani and Mori, 2007; Tuzel *et al.*, 2008), or car detection (Zheng and Liang, 2009; Cheng *et al.*, 2006). Object detection is usually a two-phase process, where binary classification is followed by feature extraction from a local image patch. Different image features (Viola and Jones, 2004; Ahonen *et al.*, 2006; Berg and Malik, 2001; Sabzmeydani and Mori, 2007; Cheng *et al.*, 2006) and binary classifiers (Freund, 2001; Breiman, 2001) are available. To our knowledge such an image processing approach has not been applied to automate peak picking in which it could provide a reliable algorithm for the identification of peaks in multidimensional NMR spectra.

In this work, we developed a peak picking algorithm based on computer vision which uses a technique called Histogram of Oriented Gradients (HOG) (Dalal and Triggs, 2005) as a feature extraction procedure and Support Vector Machine (SVM) (Cortes and Vapnik, 1995) as a binary classifier for reasons discussed later in the text. As input, the CV-Peak Picker accepts any kind of spectra with different data formats (Sparky [UCSF], Topspin [Bruker] or VNMR [Agilent (Varian)]) and returns peak lists for a given spectrum in Sparky format (http://www.cgl.ucsf.edu/home/sparky). The general scheme of the CV-Peak Picker function is illustrated in Figure 1.

2 Methods

2.1 General strategy

The following features are used in CV-Peak Picker:

- *Extrema selection*: large numbers of potential peaks are identified in the spectra.
- *Volume calculation*: 3D/4D spectra are dissected into 2D layers perpendicular to one/two axes and peak volumes calculated.
- *Bounding box*: largest and smallest peaks in 2D spectra are confined by the largest and the smallest bounding boxes.
- *Symmetrization*: 2D symmetrization of peaks affords a deconvolution of overlapped peaks.
- *Shape mapping*: peak shape is represented by Histograms of Oriented Gradients (HOG) (Dalal and Triggs, 2005) and translated from pixel image to feature descriptor.
- *Feature pyramid*: combining features of the shape extracted with bounding boxes of different sizes.
- *Peak classification*: all recognized and dispersed shapes are compared with a training set consisting of manually identified shape set that contains real peaks and peak artifacts; the comparison is performed with Support Vector Machines (SVMs).

2.2 Extrema selection

The NMR spectrum is an array of pixels having individual intensities; e.g. a 3D spectrum consists of intensity values $I_{x,y,z}$ with the Cartesian coordinates x, y, z along the three frequency axes. After exclusion of a user-defined spectral band around the water resonance the spectrum is scanned with a $3 \times 3 \times 3$ pixel cube. First local extrema are identified by the criterion that the central pixel has the single largest or smallest value of all adjacent pixels. Then, the spectrum is cut into 2D layers perpendicular to one axis, and henceforth it is analyzed layer-by-layer.

2.3 Volume calculation

For all extrema in a particular 2D layer the peak volume is defined as the sum of the intensities of data points (pixels) in small areas around the extremum starting with 3×3 pixels with the extremum in the center. The number of pixels is iteratively optimized based on the peak volume as proposed by Liu *et al.* (2012). The extrema are sorted according to their absolute volumes starting with the largest value. CV-Peak Picker analyzes N peaks per layer (default $N = 500$, any N can be chosen). N was set based on $^{13}C/^{15}N$-resolved 3D NOESY spectra which provide a high dynamic range in volumes and contain much more resonances than through-bond correlation experiments including TOCSY. N must be larger than the expected number of real peaks; larger N values increase the scanning time.

2.4 Bounding box

Peaks with absolute volumes above the cut-off value are tightly confined by the *bounding box* (Fig. 1A) which varies in size between the largest and smallest peak. Dimensions of the largest and smallest bounding boxes are predefined by the user visually within a 2D layer of the spectrum. The area of all bounding boxes is rescaled to the default size of 32×32 pixels (see Supporting Information for details). The features will be extracted within the bounding box of the individual peaks. The manual choice of the maximal and the minimal bounding box affects the result in that the size of the bounding box determines the accuracy and sensitivity of the resulting peak list. A larger box increases the chances to pick all real peaks, but also allows the algorithm to pick more artifacts.

2.5 Symmetrization

Overlapping peaks may strongly influence feature values and thus cause significant irregularities in shape mapping. We empirically found better accuracy in peak picking, when HOG features are extracted before and after symmetrization of a peak and thus, both are included into the peak descriptor (Supplementary Fig. S1A). For local symmetrization the following equation is applied:

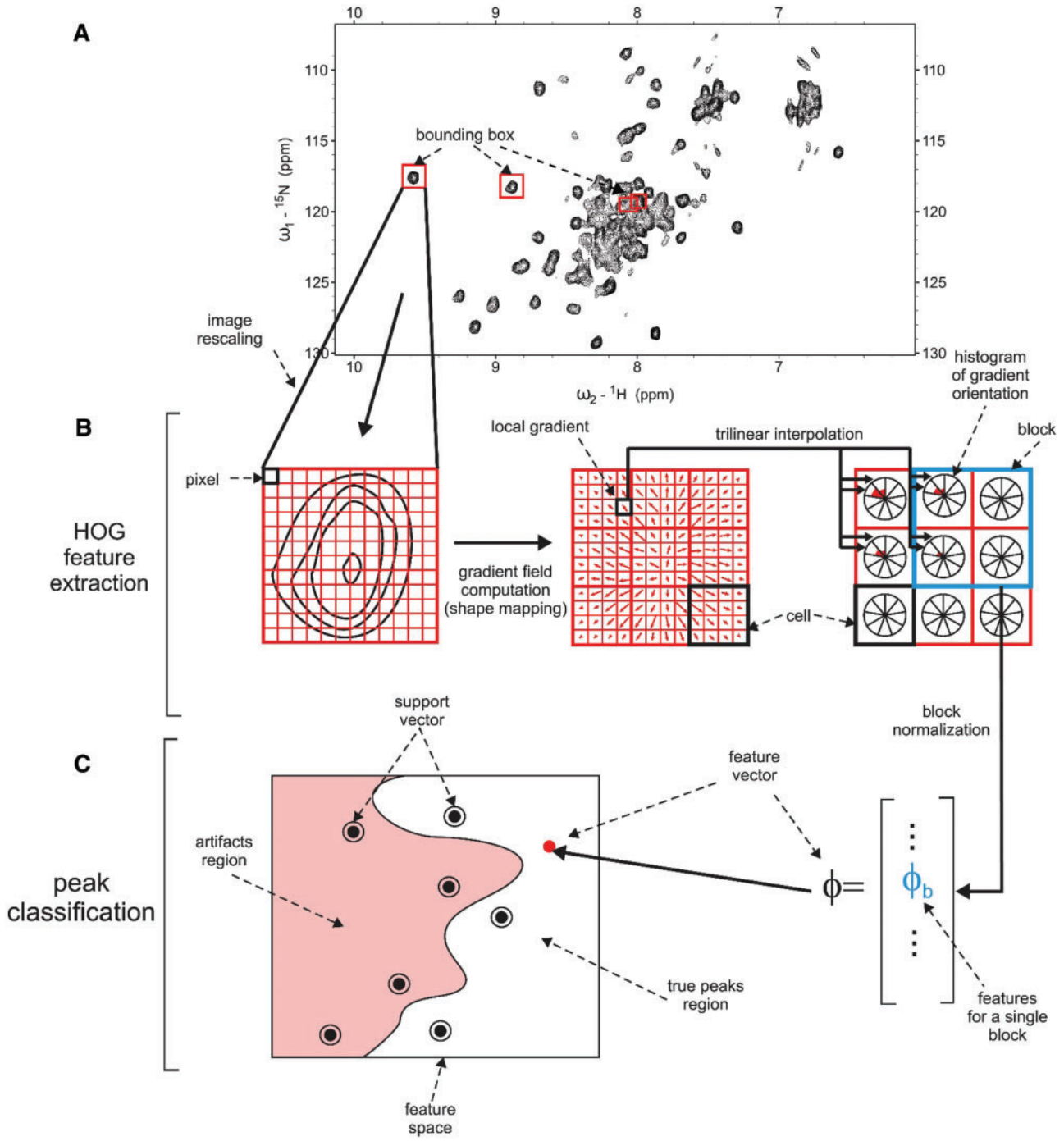

Fig. 1. Schematics showing the principle features of CV-Peak Picker. First, thousands of local extrema are identified in the 3D/4D spectrum (not shown) of interest. Then the spectrum is dissected into 2D layers and peak volumes are calculated in individual layers. (**A**) Identified peaks are confined by the bounding boxes and rescaled to the default size of 32×32 pixels. (**B**) The spectral data in the bounding box is transformed into feature space by shape mapping. A two-dimensional local gradient is assigned to each pixel using Histogram of Oriented Gradients (HOG) (Dalal and Triggs, 2005). For better stability local gradients are averaged over non-overlapping cells with 4×4 pixels by trilinear interpolation (large black squares). Next, for each cell a histogram of gradient orientation is calculated: 9 histogram bins evenly distributed over 360° are filled by adding partial gradient magnitudes calculated using trilinear interpolation, i.e. each gradient magnitude is distributed among two closest orientation bins in four adjacent cells according to Dalal and Triggs (Dalal and Triggs, 2005). Then cells are organized into blocks consisting of 2×2 cells (blue square) and bin values normalized. (**C**) The resulting bin values obtained in (B) are added into final feature vectors. These feature vectors (peak descriptors) are compared with the support vectors in the feature space. This comparison is performed using the Gaussian kernel and allows classification of the peak into the region of true peaks or peak artifacts in feature space. The flow of the algorithm is presented by solid black arrows, and elements of the flow chart are indicated with dashed black arrows. Two main layers of the algorithm, HOG feature extraction and peak classification, are restrained by the single brackets. In the program we actually use a total of 16 cells (4×4) cells each consisting of 64 (8×8) pixels. The scheme presented here is demonstrative and for clarity we use 9 cells (3×3) each consisting of 16 (4×4) pixels

$$\bar{I}_{u,v} = \begin{cases} \max\{0,\min\{I_{\lceil x\rceil-u,\lceil y\rceil-v}, I_{\lfloor x\rfloor+u,\lfloor y\rfloor+v}\}\}, & \text{if } I_{\lfloor x\rfloor,\lfloor y\rfloor} \geq 0 \\ \min\{0,\max\{I_{\lceil x\rceil-u,\lceil y\rceil-v}, I_{\lfloor x\rfloor+u,\lfloor y\rfloor+v}\}\}, & \text{if } I_{\lfloor x\rfloor,\lfloor y\rfloor} < 0 \end{cases} \quad (1)$$

where x,y denote the coordinates of the bounding box center (which can consist of non-integer values); $\lfloor\cdot\rfloor, \lceil\cdot\rceil$ denote floor and ceiling operators, respectively; the peak extremum is located in the pixel ($\lfloor x\rfloor$,$\lfloor y\rfloor$); the numbers (u,v) vary from $\lfloor -0.5M + 1\rfloor$ to $0.5M$, where M denotes the width of bounding box after rescaling.

The symmetrized spectrum is subtracted from the original spectrum and an additional round of extrema selection is performed. This provides a systematic detection of overlapped peaks.

2.6 Shape mapping

The crucial aspect in choosing the right features is their discriminative property, i.e. their values should differ significantly between real peaks and artifacts. One pronounced difference is the shape of the peaks. Real peaks usually exhibit a Lorentzian or Gaussian type line shape unlike artifacts, which have varying shapes with less symmetry. For a peak descriptor we selected 'Histogram of Oriented Gradients' (HOG) due to its flexibility in describing various irregular shapes. HOG was successfully applied in computer vision for object detection in photographs (Dalal and Triggs, 2005; Felzenszwalb *et al.*, 2010). For the *bounding box* the vector field consisting of intensity gradients is created (Fig. 1B). The horizontal gradient g_l^H for a pixel l is calculated from three consecutive pixels k, l, m with intensities I_k, I_l and I_m, respectively, by

$$g_l^H = 0.5(I_m - I_k). \quad (2)$$

The vertical gradients g_l^V are calculated correspondingly. Subsequently, vertical and horizontal gradients are combined so that the magnitude of the resulting diagonal gradient g_l, for each pixel l equals:

$$g_l = \sqrt{(g_l^H)^2 + (g_l^V)^2} \quad (3)$$

and its orientation is:

$$\theta_l = arctg\left(\frac{g_l^V}{g_l^H}\right) \quad (4)$$

For further evaluation, the gradient vector field within a bounding box is divided into 16 non-overlapping cells with 8×8 pixels each (Fig. 1B). Based on the cells the histogram of gradient orientation (HOG) is calculated. For the HOG the unit circle is divided into 9 equal parts forming the histogram bins. The bins are filled according to the gradient orientations θ_l and the corresponding partial magnitudes of the gradients are calculated by trilinear interpolation (Dalal and Triggs, 2005). This interpolation makes the feature descriptor less susceptible to small perturbations in the shape. To increase the robustness of the feature descriptor even further, the contribution of the noise must be minimized. To this end the cells are grouped into blocks consisting of 2×2 cells. Final features are a result of normalization of the histogram values in every single block as described by Dalal and Triggs (2005).

2.7 Feature pyramid

The size of the bounding box depends on the peak size in a contour plot which also represents its volume. The program uses K bounding boxes with different sizes (for details see Supporting Information). From the spectral information within individual bounding boxes we extract features that we combine with a procedure called feature pyramid (Lowe, 2004). Formally, the feature pyramid Φ_i is defined by

$$\Phi_i = (\boldsymbol{\phi}_{i,1}, \ldots, \boldsymbol{\phi}_{i,K}) \quad (5)$$

where $\boldsymbol{\phi}_{i,k}$ is a vector of extracted features for the peak (x_i, y_i) using the k-th area, where $k = 1, \ldots, K$, for which we calculate exactly the same features, i.e. HOG with and without symmetrization. To confirm the reproducibility of the feature extraction procedure, the peak previously confined by the bounding box, is rescaled to default 32×32 pixels (for details see Supporting Information).

2.8 Peak classification

From every feature pyramid $\boldsymbol{\Phi}_i$ a sequence of real valued responses $\{r_{i,1}, \ldots, r_{i,K}\}$ can be extracted, in which larger $r_{i,k}$ values represent stronger evidence that the peak is in the positive class (real peak), the value of p_i is set to the strongest response

$$p_i = \begin{cases} 1 & if \quad \max_k r_{i,k} \geq r_0 \\ 0 & if \quad \max_k r_{i,k} < r_0 \end{cases}, \quad (6)$$

where r_0 is a constant threshold tuned by classifying a preliminary referenced set of peaks (see Supplementary Fig. S1B). The classifier is trained on a set of true peaks, where the correct size of the bounding box was set manually. Consequently, it is highly probable that a real peak with an improper bounding box size will be classified as an artifact, because it is divergent from the real peaks in the training set. However, at least one of the different bounding boxes used between the manually defined minimum and maximum, will be classified as highly positive.

To calculate the $r_{i,k}$ a classifier known as Support Vector Machine (SVM) is used (Cortes and Vapnik, 1995). SVMs are currently one of the basic techniques in window-based object detection where binary classification is used [along with AdaBoost and Random Forests (Breiman, 2001; Freund *et al.*, 1999)]. A major advantage is that they are fast and easily trained to obtain a robust classifier. An SVM classifier has the following form

$$r_{i,k} = \sum_j \alpha_j K(\boldsymbol{\phi}_{i,k}, \boldsymbol{\phi}^{(j)}) + \alpha_0, \quad (7)$$

where α_j denotes model parameters, $K(\cdot,\cdot)$ is defined in Equation (8), and $\boldsymbol{\phi}^{(j)}$ are the reference descriptors, also called support vectors. Model parameters and support vectors are obtained during the SVM learning process based on a set of examples $\left\{\left(\boldsymbol{\phi}^{(1)}, p^{(1)}\right), \ldots, \left(\boldsymbol{\phi}^{(N)}, p^{(N)}\right)\right\}$, the area of which is called here the training set (Cortes and Vapnik, 1995).

Our training set consists of 7566 manually selected example peaks which cover various shapes including both real peaks (3887) and artifacts (3679). The set containing 13 spectra was created from HNCO, HNCA, HSQC, CBCA(CO)NH, HNCACB experiments measured with proteins that contained between 114 and 209 residues.

The peak descriptor (feature vector) is compared with support vectors (Fig. 1C) using the Gaussian kernel

$$K(\boldsymbol{\phi}, \boldsymbol{\phi}') = \exp(-\gamma\|\boldsymbol{\phi} - \boldsymbol{\phi}'\|^2). \quad (8)$$

With increasing similarity of the feature vectors the kernel approaches its maximum value of 1. Thereby, the support vectors that are similar to the feature vector, have higher impact on the response. Thus, it is important to have a representative training set so that the support vectors cover diverse regions in the feature space.

The precision parameter γ is tuned according to the standard procedures (see Supporting Information).

After the peak classification process the positively classified peaks from all layers are entered into a three-dimensional peak list. This list is imported into the program Sparky, the peak extrema are refined by quadratic interpolation and the list is stored in Sparky format.

3 Results

We evaluated the CV-Peak Picker algorithm by comparing the results with those of established programs PICKY (Alipanahi *et al.*, 2009), WaVPeak (Liu *et al.*, 2012) using a set of benchmark spectra and additional spectra (Table 1). Spectra of only 3 proteins from the original benchmark set were made available to us by the authors of PICKY and WaVPeak. We also tested our CV based algorithm on a set of different spectra obtained for proteins with molecular weights from 13 to 130 kDa (Table 2) to test the high potential of the CV software. All spectra were picked independently of any other spectrum or input, i.e. the number of artifacts will be reduced when combining peak lists from different spectra which is necessary for sequence specific assignments or structure calculations. For most spectra all real peaks were correctly picked with a remarkably small number of artifacts. In the evaluation of our results we used generally accepted statistical measures: precision, recall and F-measure. Precision, P, and recall, R, are defined as $P = (TP/(TP + FP))*100$ and $R = (TP/(TP + FN))*100$, where TP stands for true positives ('real peaks' which were classified as 'true peaks'), FN stands for false negatives ('real peaks' which were classified as 'artifacts') and FP stands for false positives ('false peaks' classified as 'real peaks'). Combination of these two measures is represented as F-measure, $F = 2RP/(R + P)$, and allows for direct comparisons of performance between different methods (Powers, 2011). Values of the F-measure fall into the range from 0 to 100, where 100 represents highest performance of the method.

Table 1. Comparison between the performance of the CV-Peak Picker, PICKY and WaVPeak[a,b,c,d,e]

Protein name	Spectrum			CV-Peak Picker			PICKY 2009						WaVPeak 2012					
	Experiment	SNR[d]	No.of real peaks	F-measure	Precision	Recall	F-measure		Precision		Recall		F-measure		Precision		Recall	
COILIN	CBCA(CO)NH	40	289	73	70	76	66	-7	54	-16	86	10	60	-13	55	-15	66	-10
	HNCO	71	165	81	77	86	73	-9	58	-19	97	11	68	-13	62	-15	75	-11
	HNCACB	22	317	85	87	83	64	-21	54	-33	78	-5	62	-23	57	-30	68	-15
	HSQC	45	111	97	96	98	81	-16	70	-26	97	-1	82	-15	75	-21	90	-8
VRAR	CBCA(CO)NH	25	111	90	85	95	77	-13	71	-14	83	-12	74	-15	68	-17	82	-13
	HNCO	56	58	98	99	98	86	-12	84	-15	89	-9	85	-14	78	-21	93	-5
	HNCACB	23	174	91	96	87	70	-21	72	-24	69	-18	62	-29	57	-39	68	-19
	HSQC	68	57	98	99	98	90	-9	87	-12	93	-5	88	-10	81	-18	97	-1
HACS1	CBCA(CO)NH	53	167	93	97	90	74	-19	61	-36	94	4	81	-13	74	-23	89	-1
	HNCO	86	89	94	96	93	75	-20	62	-34	94	1	84	-10	77	-19	93	0
	HNCACB	34	240	93	93	94	64	-30	52	-41	82	-12	77	-16	71	-22	85	-9
	HSQC	51	87	94	91	98	79	-16	67	-24	95	-3	88	-6	81	-10	97	-1
pRN1	HSQC	61	201	96	96	96	94	-2	90	-6	98	2	86	-10	77	-19	97	1
KcsA	HN(CO)CA	13	96	77	81	73	49	-28	46	-35	53	-20	42	-35	29	-52	70	-3
FimAwt	HN(CO)CA	49	177	94	95	92	92	-2	89	-6	94	2	93	-1	92	-3	95	3
Nlgn-3	HNCACB	24	482	75	79	72	75	0	71	-8	79	7	33	-42	27	-52	40	-32
TM1290	HNCA	176	237	94	98	91	92	-2	92	-6	91	0	88	-6	82	-16	95	4
Statistics		Mean		90	90	89	76	-13	69	-21	87	-3	74	-16	67	-23	82	-7
		Correlation between F-measure and SNR[e]		0,37			0,57						0,53					

[a]The benchmark spectra were provided by Prof. Xin Gao, King Abdullah University of Science and Technology, Thuwal, Saudi Arabia and Prof. Ming Li, University of Waterloo, Waterloo, ON, Canada. The proteins COILIN, VRAR and HACS1 comprise 98, 72 and 74 amino acids, respectively.

[b]Performance measures, Recall (R) and Precision (P), are defined in reference (Alipanahi *et al.*, 2009); $R = TP/(TP + FN)*100$ and $P = TP/(TP + FP)*100$, where TP stands for true positives ('real peaks' which were classified as 'true peaks'), FN stands for false negatives ('real peaks' which were classified as 'artifacts') and FP stands for false positives ('false peaks' classified as 'real peaks').

[c]F-measure is a value that combines precision and recall using harmonic mean. It ranges from 0 to 100, where 100 is the score for best performance. The last column contains the difference of the F-measure of CV-Peak Picker and the higher F-measure value of two programs, PICKY and WaVPeakThe columns F-measure, Precision and Recall contain double values for PICKY and WaVPeak. Colored values show difference of the parameter (F-measure, Precision and Recall) vs. value of this measure for CV-Peak Picker. Values colored in red show worse performance of the measure and in green better performance compared to CV-Peak Picker.

[d]Signal-to-noise (SNR) was calculated by dividing the mean of the real signals in the spectrum (excluding water signal) by the standard deviation of the intensity of the noise.

[e]The correlation coefficient was calculated with the standard function 'Correlation Coefficient' in Microsoft Excel.

Table 2. Summary of the performance of the CV-Peak Picker on a set of selected NMR spectra[a]

Protein	Protein concentration (mM)	Spectrometer frequency (MHz)	Probehead	Spectrum	F-measure	Precision	Recall
pRN1 primase-polymerase (Lipps *et al.*, 2004)	0.8	700	cryoprobe	HSQC	96	96	96
KcsA (Kent *et al.*, 2007)	0.7	700	cryoprobe	HNCOCA	77	81	73
FimAwt (Walczak *et al.*, 2014)	1.2	750	room temp.	HNCOCA	94	95	92
Nlgn-3 (Wood *et al.*, 2012)	1.0	600	room temp.	HNCACB	75	79	72
TM1290 (Etezady-Esfarjani *et al.*, 2003)	1.5	750	room temp.	HNCA	94	98	91

[a]Proteins pRN1 primase-polymerase, FimA, Nlgn-3 (intrinsically disordered) and TM1290 comprise 209, 159, 127 and 116 amino acids, respectively. KcsA is a homo-tetrameric protein with160 residues per subunit solubilized in detergent micelles.

Table 3. Complete list of CV-Peak Picker parameters that are specified by the user

Parameter	Description	Specification
Size of the scanning window	**Obligatory**: Estimation of the proper size of the scanning window is crucial for the scanning procedure. The peak picker performs well if the side length of the window selected by the user varies up to ± 50% from the optimal one.	CV-Peak Picker offers an interactive tool which allows a user to draw a scanning window on the contour plot of the spectrum.
Threshold r_0	**Optional**: The parameter can be used for fine tuning after all peaks are evaluated using SVM. Often change of the default value is not necessary.	The parameter can be adjusted using an interactive graphical user interface, which updates visualization of the scanning results in real time according to the value of r_0.
Number of scanned peaks per layer	**Optional**: The parameter is introduced exclusively for performance optimization. A user can request to scan all peaks in the spectrum (even small artifacts) by assigning an arbitrarily large value to this parameter. Nevertheless, scanning the biggest 500 peaks on each layer of a triple-resonance spectrum is usually sufficient.	The parameter can be specified in the CV-Peak Picker configuration file.

Table 1 compares the peak picking performance of CV-Peak Picker, PICKY and WaVPeak on the set of benchmarked spectra as introduced by (Alipanahi *et al.*, 2009) and on more challenging spectra selected by us. A comparison of the F-measures between CV-Peak Picker, PICKY and WaVPeak shows that CV-Peak Picker performs substantially better on each of the benchmark spectra. In the same table correlation coefficient between F-measure and signal-to-noise ratio is presented.

In Table 2, we summarize peak picking results for spectra which are more challenging than the benchmark spectra. In two cases for the highly overlapped spectra of the 130 kDa helical membrane protein KcsA alpha and of intrinsically disordered Nlgn-3, CV-Peak Picker reaches F-measures of about 75. For the remaining proteins high F-measure scores of approximately 95 are achieved. Exemplary planes of these five spectra are shown in Supplementary Figure S2. The reference peak lists for the spectra in Table 2, were obtained from original peak lists (provided by the authors or found in assignment papers) which we manually corrected as they contained only assigned peaks and not all real peaks. The performance of CV-Peak Picker is exemplified in Supplementary Figure S3 with six consecutive cross sections of the HN(CO)CA spectrum of KcsA. Table 3 lists the 3 parameters (1 obligatory and 2 optional) which must be set before running the CV-Peak Picker.

In the Supporting Information, we graphically represent the dependence of the performance (F-measure) of CV-Peak Picker on the selection of the size of bounding box (Supplementary Fig. S4) and on the threshold value r_0 (Supplementary Fig. S5). These figures show that the values for the bounding box and r_0 can be chosen in a wide range without substantially changing the resulting peak list. Further, in Supplementary Table S2, we provide average scanning times per 2D layer of various 3D heteronuclear correlation spectra measured with different proteins.

4 Discussion

The application of computer vision techniques for a peak picking algorithm applied to NMR spectra described here, results in fully automated peak recognition in a wide range of triple-resonance spectra obtained with five different globular proteins with molecular weights between 13 and 26 kDa and a uniformly deuterated tetrameric (4×17.5 kDa) alpha helical membrane protein in detergent micelles (130 kDa in complex). Moreover it outperforms other currently used peak pickers as evaluated by benchmark spectra. Features and system requirements of the most popular peak picking programs are presented in Supplementary Table S1.

A standard measure used for comparison of the performance of peak picking algorithms, the F-measure, is a combination of two components: recall and precision. The precision measure contains information on the excessively picked artifacts, while recall contains information on missed real peaks. In practice, for structural studies recall has significantly higher importance than precision as loss of information (missed real peaks) often cannot be compensated and thus it may lead to incomplete assignments. Conversely, excessively picked artifacts are rather easily rejected in an assignment process. In Table 1 the performance of CV-Peak Picker and WaVPeak as well as PICKY are compared. CV-Peak Picker always has superior precision scores and overall F-measures. However, in ten cases CV-Peak Picker reaches slightly worse recall scores than PICKY and/or WaVPeak. Although these differences are relatively small, we

speculate that a substantial increase of the number of peaks in the CV-Peak Picker's training set, might help reaching higher recall scores. On the other hand, WaVPeak requires a preset number of the expected real peaks in the spectrum, which results in a classification of questionable peaks as real. Table 1 also presents a correlation between the F-measure and the signal-to-noise ratio (SNR). CV-Peak Picker reaches the similar F-measure at low SNR which is a substantial advantage for its versatility and usefulness in real spectra with suboptimal SNR.

The applications of CV-Peak Picker in this work demonstrate that the analysis of NMR spectra of proteins with difficult, highly overlapped spectra including IDPs or molten globules (Walczak *et al.*, 2014) can be fully automated. User attention can be focused on validation of the results. We believe that highest accuracy can be achieved by coupling our software with powerful assignment programs as e.g. FLYA (López-Méndez and Güntert, 2006). The methodology presented in this paper has high potential for investigations and analysis of NMR spectra of nucleic acids and carbohydrates, and could be adapted for peak picking of solid state NMR spectra.

Acknowledgements

We gratefully acknowledge Dr. Gerrit Daubner, Dr. Pierre Barraud, Dr. Antoine Cléry, Dr. Benjamin I. Leach, Prof. John H. Bushweller, Prof. Frans A. Mulder, Prof. Roland Riek and Prof. Frederic Allain for providing us with NMR spectra which were used for the development and the evaluation of the algorithm. We also thank Prof. Ming Li (University of Waterloo, Canada) and Prof. Xin Gao (King Abdullah University of Science and Technology, Thuwal, Saudi Arabia) for providing benchmark spectra. We are grateful to Dr. Fionna Loughlin for the careful reading of the manuscript.

Funding

Funding for this research was provided by the Ministry of Science and Higher Education Grant B40235/K3/W8 (AG), Poland, and by ETH Research Grant ETH-23 10-2 (MJW), Switzerland.

Conflict of Interest: none declared.

References

Abbas,A. *et al.* (2013) Automatic peak selection by a Benjamini–Hochberg-based algorithm. *PLoS One*, **8**, e53112.

Ahonen,T. *et al.* (2006) Face description with local binary patterns: Application to face recognition. *IEEE Trans. Pattern Anal. Mach Intellig*, **28**, 2037–2041.

Alipanahi,B. *et al.* (2009) PICKY: a novel SVD-based NMR spectra peak picking method. *Bioinformatics*, **25**, i268–i275.

Antz,C. *et al.* (1995) A general Bayesian method for an automated signal class recognition in 2D NMR spectra combined with a multivariate discriminant analysis. *J. Biomol. NMR*, **5**, 287–296.

Banci,L. *et al.* (2010) NMR in structural proteomics and beyond. *Prog. Nucl. Magn. Reson. Spectrosc.*, **56**, 247–266.

Baran,M.C. *et al.* (2004) Automated analysis of protein NMR assignments and structures. *Chem. Rev.*, **104**, 3541–3556.

Berg,A.C. and Malik,J. (2001) Geometric blur for template matching. In: *Proc 2001 IEEE Comp Soc Conf IEEE CVPR 2001*, p. I-607–I-614, vol. 601.

Breiman,L. (2001) Random forests. *Mach. Learn.* **45**, 5–32.

Carrara,E.A. *et al.* (1993) Neural networks for the peak-picking of nuclear magnetic resonance spectra. *Neural Netw.* **6**, 1023–1032.

Cheng,H. *et al.* (2006) Boosted Gabor features applied to vehicle detection. In: *18th Internat Conf Pattern Recogn 2006. ICPR 2006*, IEEE, p. 662–666.

Cheng,Y. *et al.* (2014) Bayesian peak picking for NMR spectra. *Genomics Proteomics Bioinf.*, **12**, 39–47.

Cieslar,C. *et al.* (1988) Computer-aided sequential assignment of protein 1H NMR spectra. *J. Magn. Reson. (1969)*, **80**, 119–127.

Coles,M. *et al.* (2003) NMR-based screening technologies. *Drug Discov. Today*, **8**, 803–810.

Cortes,C. and Vapnik,V. (1995) Support-vector networks. *Mach. Learn.*, **20**, 273–297.

Dalal,N. and Triggs,B. (2005) Histograms of oriented gradients for human detection. In: *IEEE Comp Soc Conf IEEE CVPR 2005*, p. 886–893.

Etezady-Esfarjani,T. *et al.* (2003) NMR assignment of the conserved hypothetical protein TM1290 of Thermotoga maritima. *J. Biomol. NMR*, **25**, 167–168.

Felzenszwalb,P.F. *et al.* (2010) Object detection with discriminatively trained part-based models. *IEEE Trans. Pattern Anal. Mach. Intell.*, **32**, 1627–1645.

Freund,Y. (2001) An adaptive version of the boost by majority algorithm. *Mach. Learn.*, **43**, 293–318.

Freund,Y. *et al.* (1999) A short introduction to boosting. *J. Jpn. Soc. Artif. Intell.*, **14**, 1612.

Garrett,D.S. *et al.* (1991) A common sense approach to peak picking in two-, three-, and four-dimensional spectra using automatic computer analysis of contour diagrams. *J. Magn. Reson. (1969)*, **95**, 214–220.

Güntert,P. (2004) Automated NMR structure calculation with CYANA. *Methods Mol. Biol.*, **278**, 353–378.

Hajduk,P.J. *et al.* (1999) NMR-based screening in drug discovery. *Q. Rev. Biophys.*, **32**, 211–240.

Herrmann,T. *et al.* (2002) Protein NMR structure determination with automated NOE-identification in the NOESY spectra using the new software ATNOS. *J. Biomol. NMR.*, **24**, 171–189.

Herrmann,T. *et al.* (2002) Protein NMR structure determination with automated NOE assignment using the new software CANDID and the torsion angle dynamics algorithm DYANA. *J. Mol. Biol.*, **319**, 209–227.

Hiller,S. *et al.* (2008) APSY-NMR with proteins: practical aspects and backbone assignment. *J. Biomol. NMR*, **42**, 179–195.

Jung,Y.-S. and Zweckstetter,M. (2004) Mars—robust automatic backbone assignment of proteins. *J. Biomol. NMR*, **30**, 11–23.

Kent,A.B. *et al.* (2007) Conformational dynamics of the KcsA potassium channel governs gating properties. *Nat. Struct. Mol. Biol.*, **14**, 1089–1095.

Kleywegt,G.J. *et al.* (1990) A versatile approach toward the partially automatic recognition of cross peaks in 2D 1 H NMR spectra. *J. Magn. Reson. (1969)*, **88**, 601–608.

Koradi,R. *et al.* (1998) Automated peak picking and peak integration in macromolecular NMR spectra using AUTOPSY. *J. Magn. Reson.*, **135**, 288–297.

Lipps,G. *et al.* (2004) Structure of a bifunctional DNA primase-polymerase. *Nat. Struct. Mol. Biol.*, **11**, 157–162.

Liu,Z. *et al.* (2012) WaVPeak: picking NMR peaks through wavelet-based smoothing and volume-based filtering. *Bioinformatics*, **28**, 914–920.

Lowe,D.G. (2004) Distinctive image features from scale-invariant keypoints. *Int. J. Comp. Vis.*, **60**, 91–110.

López-Méndez,B. and Güntert,P. (2006) Automated protein structure determination from NMR spectra. *J. Am. Chem. Soc.*, **128**, 13112–13122.

Parsons,L. and Orban,J. (2004) Structural genomics and the metabolome: combining computational and NMR methods to identify target ligands. *Curr. Opin. Drug Discovery Dev.*, **7**, 62–68.

Pellecchia,M. *et al.* (2004) NMR-based techniques in the hit identification and optimisation processes. *Expert. Opin. Therap. Targ.*, **8**, 597–611.

Pellecchia,M. *et al.* (2008) Perspectives on NMR in drug discovery: a technique comes of age. *Nat. Rev. Drug Discov.*, **7**, 738–745.

Powers,D.M. (2011) Evaluation: from precision, recall and F-measure to ROC, informedness, markedness & correlation. *JMLT*, **2**, 37–63.

Sabzmeydani,P. and Mori,G. (2007) Detecting pedestrians by learning shapelet features. In: IEEE Conference on CVPR'07. IEEE, p. 1–8.

Tikole,S. *et al.* (2014) Peak picking NMR spectral data using non-negative matrix factorization. *BMC Bioinf.*, **15**, 46.

Tuzel,O. *et al.* (2008) Pedestrian detection via classification on riemannian manifolds. *IEEE Trans. Pattern Anal. Mach. Intell.*, **30**, 1713–1727.

Viola,P. and Jones,M.J. (2004) Robust real-time face detection. *Int. J. Comp. Vis.*, **57**, 137–154.

Walczak,M.J. *et al*. (2014) Intramolecular donor strand complementation in the *E. coli* type 1 pilus subunit FimA explains the existence of FimA monomers as off-pathway products of pilus assembly that inhibit host cell apoptosis. *J. Mol. Biol.*, **426**, 542–549.

Walczak,M.J. *et al*. (2014) The RING domain of the Scaffold protein Ste5 adopts a molten globular character with high thermal and chemical stability. *Angew. Chem. Int. Ed. Engl.*, **53**, 1320–1323.

Wood,K. *et al*. (2012) Backbone and side chain NMR assignments for the intrinsically disordered cytoplasmic domain of human neuroligin-3. *Biomol. NMR Assign.*, **6**, 15–18.

Yee,A. *et al*. (2006) Solution NMR in structural genomics. *Curr. Opin. Struct. Biol.*, **16**, 611–617.

Zheng,W. and Liang,L. (2009) Fast car detection using image strip features. In: IEEE Conference on CVPR 2009. IEEE, p. 2703–2710.