

Automated Discrimination of Pathological Regions in Tissue Images: Unsupervised Clustering vs Supervised SVM Classification

*Original*

Automated Discrimination of Pathological Regions in Tissue Images: Unsupervised Clustering vs Supervised SVM Classification / DI CATALDO, Santa; Ficarra, Elisa; Macii, Enrico - In: Biomedical Engineering Systems and Technologies HEIDELBERG : Springer, 2008. - ISBN 9783540922186. - pp. 344-356 [10.1007/978-3-540-92219-3\_26]

*Availability:*

This version is available at: 11583/1829621 since: 2019-04-19T12:04:40Z

*Publisher:*

Springer

*Published*

DOI:10.1007/978-3-540-92219-3\_26

*Terms of use:*

This article is made available under terms and conditions as specified in the corresponding bibliographic description in the repository

*Publisher copyright*

(Article begins on next page)

# Automated Discrimination of Pathological Regions in Tissue Images: Unsupervised Clustering vs Supervised SVM Classification

Santa Di Cataldo, Elisa Ficarra, and Enrico Macii

Dep. of Control and Computer Engineering, Politecnico di Torino, Cso Duca degli Abruzzi  
24, 10129 Torino, Italy  
{santa.dicataldo, elisa.ficarra, enrico.macii}@polito.it

**Abstract.** Recognizing and isolating cancerous cells from non pathological tissue areas (e.g. connective stroma) is crucial for fast and objective immunohistochemical analysis of tissue images. This operation allows the further application of fully-automated techniques for quantitative evaluation of protein activity, since it avoids the necessity of a preventive manual selection of the representative pathological areas in the image, as well as of taking pictures only in the pure-cancerous portions of the tissue. In this paper we present a fully-automated method based on unsupervised clustering that performs tissue segmentations highly comparable with those provided by a skilled operator, achieving on average an accuracy of 90%. Experimental results on a heterogeneous dataset of immunohistochemical lung cancer tissue images demonstrate that our proposed unsupervised approach overcomes the accuracy of a theoretically superior supervised method such as Support Vector Machine (SVM) by 8%.

**Keywords:** Tissue segmentation, tissue confocal images, immunohistochemistry, K-means clustering, Support Vector Machine.

## 1 Introduction

Detecting tumor areas in cancer tissue images and disregarding non pathological portions such as connective tissue are critical tasks for the analysis of disease state and dynamics. In fact, by monitoring the activity of proteins involved in the genesis and the development of multi-factorial genetic pathologies we can obtain a useful diagnostic tool. It leads to classify the pathology in a more accurate way through its particular genetic alterations, and to create new opportunities for early diagnosis and personalized predictive therapies [1].

An approach for monitoring and quantifying the protein activity in pathological tissues is to analyze, for example, images of the tissue where the localization of proteins is highlighted by fluorescent marked antibodies that can detect and link the target proteins. The antibodies are marked with particular stains whose intensity is related to protein activity intensity. This procedure is called *immunohistochemistry* (IHC). The increased use of immunohistochemistry in both clinical and basic research settings has led to the development of techniques for acquiring quantitative

information from immunostains and automated imaging methods have been developed in an attempt to standardize IHC analysis.

Tissue segmentation for tumor areas detection is the first fundamental step of automated IHC image processing and protein activity evaluation. In fact the quantification of a target protein's activity should be performed on tumor portions of the tissue without taking into account the non pathological areas eventually present in the same IHC images. In Fig. 1 examples of IHC tissue images are reported where connective tissue (i.e. non tumoral tissue) is outlined in black (for details about these images see Section 2).

Several methods have been proposed in the last few years to perform automated segmentation of tissue images [2], [3], [4], [5], [6]. However the most accurate approaches are those that provide a well-suited framework for incorporating primary expert knowledge into the adaptation of algorithms, such as supervised learning algorithm (e.g. Neural Networks, Machine Learning, kernel-based) [6].

The most prominent algorithm among these is the *Support Vector Machine* (SVM) proposed by V.Vapnik [7] for binary classification.

SVM is a theoretically superior machine learning method which has often been shown to achieve great classification performance compared to other learning algorithms across most application fields and tasks, including image processing and tissue image processing in particular [8], [9], [10].

Moreover, the SVM method is more able to handle very high dimensional feature spaces than traditional learning approaches [11], [12]. This is in fact the case of the images targeted by our work.

However, the IHC tissue images we considered in our study present an intrinsic complexity, such as very different characteristics of staining, intensity distribution, considerable variation of tissue shape and/or size and/or orientation and, finally, considerable variation of the signal intensity within the same tissue areas due for example to superimposed staining.

Because of the heterogeneity of the representative features related to each tissue, it is very difficult for the supervised methods to obtain a satisfactory fixed classifier able to distinguish between tumor areas (i.e. epithelial tissue) and non cancerous tissue portions (such as connective tissue).

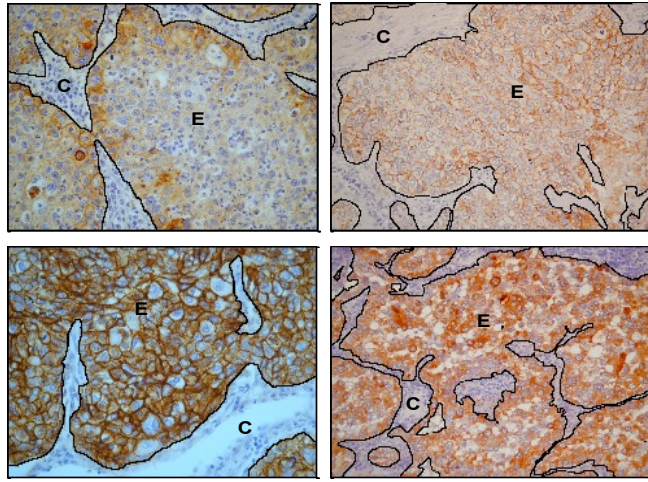
For this reason we designed a fully-automated unsupervised approach that is based only on the characteristics of the input image rather than on a fixed model of the ground truth.

In this paper we present our fully-automated unsupervised method and we compare its performance to that provided by a SVM approach applied on the same IHC tissue image target. We demonstrate that our method enables more accurate tissue segmentation compared with SVM.

In Section 2 we detail our fully-automated unsupervised method and we briefly introduce the SVM method. The implementation and the set-up are discussed in Section 3. Experimental results conducted on a large set of heterogeneous immunohistochemical lung cancer images are reported and discussed in Section 4. Finally, the Conclusions are reported in Section 5.

## 2 Method

The images we analyzed in this work were acquired through high-resolution confocal microscopy and show lung cancer tissue cells stained with marked antibodies (see Fig. 1). They are characterized by a blue hematoxylin stain as a background colour and a brown DAB stain in cellular regions where a receptor of the EGF-R/*erb-B* or TGF-alpha family is detected (i.e. membranes or cytoplasm, respectively). Cellular nuclei are blue-coloured and show a staining intensity darker than background. In all the images a remarkable portion of connective or other non cancerous tissue components is present, which appears as a blue-coloured mass (since brown DAB-stained cells are only in cancerous tissue) with quite well-defined borders.



**Fig. 1.** IHC tissue images. First row, from the left: x400 image with EGF-R positive reactions, x200 image with EGF-R positive reactions; second row: x400 image with EGF-R positive reactions, x200 image with TGF-alpha positive reactions. Connective tissue is outlined in black and labelled with *C*; epithelial tissue is labelled with *E*

Connective tissue is usually characterized by shorter inter-cellular distances and smaller nuclei than epithelial component; however, a generalization of this remark is impossible because shape and dimensions distributions of cancerous cells are often not predictable. As we outlined in the Introduction, in order to perform accurate and robust cell segmentation and protein activity quantification [13] these non cancerous tissue portions have to be identified and isolated from the representative epithelial tissue.

Here we present two different segmentation approaches to perform this critical task: i) an unsupervised procedure based on a K-means clustering of brown intensities followed by some morphological and edge-based refinement steps (see Fig. 2); ii) a supervised classification of RGB features through Support Vector Machine (see Fig. 5). Experimental results obtained with each approach on the same real-life datasets are presented and compared in Section 4.

## 2.1 Unsupervised Procedure

Since non cancerous cells do not show positive reactions at the EGF-R/TGF-alpha receptors, the monochromatic pure-DAB component instead of the original RGB image can be analyzed to perform tissue segmentation: in fact in this simpler color space connective components can be easily identified as wide bright regions with a quite homogeneous appearance (see Fig. 3(b)).

An unsupervised learning algorithm (K-means, in our work) can be efficaciously applied to isolate bright regions; then areas which show morphological and edge characteristics which are typical of connective tissue can be selected to refine tissue segmentation.



Fig. 2. Flow-chart of the unsupervised procedure based on K-means clustering.

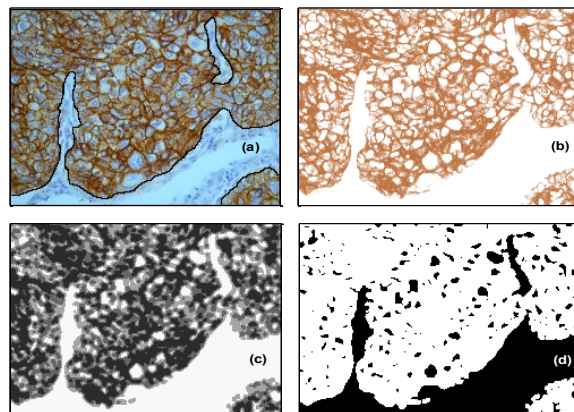


Fig. 3. Unsupervised procedure: (a) original IHC image with connective regions manually outlined (*in black*); (b) pure-DAB image (c) results after K-means clustering (pixels belonging to different clusters are mapped with grey intensity proportional to the cluster centroid); (d) cluster with highest centroid value (*in black*); as outlined in section 2.1 step 4, some small and round-shaped epithelial particles still have to be removed.

Main steps of the procedure are (see Fig. 2 and 3):

**Step 1: DAB-Component Separation.** To separate pure-DAB from pure-hematoxylin component a color deconvolution algorithm based on stain-specific RGB absorption is applied on the original RGB image [14], [15].

Differently from classical color segmentation approaches based on transformation of RGB information to HSI or to another specific color representation [16], this

method has been demonstrated to perform a good color separation even with colocalized stains.

This critical condition, due to chemical reactions of stains linking the target proteins and to the tissue superposition during the slicing of samples before image acquisition, is very common in the images targeted by our method.

For this step, the free color deconvolution plugin developed by [17] was integrated to our algorithm.

**Step 2: Preprocessing.** In pure-DAB images, connective tissue can be differentiated from epithelial tissue through its higher intensity (see Fig. 3(b)); anyway some preprocessing is needed in order to homogenize and separate the intensity distributions of the two tissues, thus improving K-means' performance.

First of all, a mean filter is performed: this operation replaces each pixel value with the average value in its neighbourhood, thus smoothing intensity peaks and decreasing the influence of single non-representative pixels.

Then a minimum filter is applied. The filter replaces pixels values with the minimum intensity values in their neighbourhood: this transformation reduces the intensity dynamic and performs a further separation of connective and epithelial intensity distributions, since the former shows minimum values higher than the latter.

**Step 3: K-Means Clustering.** To isolate bright pixels belonging to connective tissue a K-means clustering, the well-known unsupervised learning algorithm [18] which iteratively partitions a given dataset into a fixed number of clusters, is applied. This iterative partitioning minimizes the sum, over all clusters, of the within-cluster sums of point-to-cluster-centroid distances. Thus the procedure minimizes the so-called *objective function*,  $J$  in Equation 1, where  $k$  is the number of clusters,  $n$  is the number of data points and the quadratic expression is the distance measure between a data point  $x_i^{(j)}$  and the current cluster centroid  $c_j$ .

$$J = \sum_{j=1}^k \sum_{i=1}^n \|x_i^{(j)} - c_j\|^2. \quad (1)$$

The cluster with the highest centroid value is selected as representative of the connective tissue (see Fig. 3(c)). The number of clusters  $k$  was empirically set to four (see Section 3.1 for details about the parameter set-up).

**Step 4: Refinement by Size and Circularity Analysis.** Bright epithelial regions with low EGF-R/TGF-alpha activity have to be removed from the connective cluster to refine tissue segmentation.

As shown in Fig. 3(d), a large number of these regions are approximately round-shaped and are considerably smaller than connective mass: then a selective removal of particles with a low area and a high circularity compared to threshold values  $T_S$  and  $T_C$  is performed (parameters set-up in Section 3.1).

Equation 2 shows the proposed index for circularity evaluation (a value of 1 indicates a perfect circle, a value approaching 0 an increasingly elongated polygon).

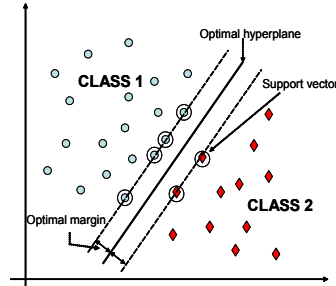
$$Circularity = 4\pi \frac{Area}{Perimeter^2}. \quad (2)$$

**Step 5: Refinement by Gradient Magnitude Analysis.** Other bright epithelial regions can be removed from the connective cluster through their edge characteristics, since connective tissue usually shows a well-defined boundary w.r.t. epithelial background in terms of intensity gradient variation. On the base of this remark, in this step areas which show along their boundary a percentage of edge pixels (i.e. pixels with high gradient intensity variation w.r.t. background) lower than a threshold value  $T_E$  are selectively removed from connective cluster (parameter set-up in Section 3.1). Edge detection is performed through a Sobel detector followed by automated intensity global thresholding.

## 2.2 Supervised Procedure

An alternate approach for tissue segmentation is supervised learning; for this purpose a *Support Vector Machine* (SVM) classification is proposed.

The SVM [7] is a theoretically superior machine learning method which has often been shown to achieve great classification performance compared to other learning algorithms across most application fields and tasks including image processing [19].



**Fig. 4.** Maximum-margin hyperplane in SVMs (linearly separable case). The boundary training instances (*support vectors*) are indicated by an extra circle.

Here we propose a procedure based on binary SVM classification, in which the input elements (in this work, small tissue regions) are associated to one of two different classes, connective or epithelial, on the base of a set of representative characteristics, the *features vector*. To perform a reliable classification, the SVM is previously trained with a set of elements whose class is well-known, the so-called *training instances*. The classification is based on the implicit mapping of data to a higher dimensional space via a *kernel function* and on the consequent solving of an optimization problem to identify the *maximum-margin hyperplane* that separates the given training instances (see Fig. 4). This hyperplane is calculated on the base of boundary training instances (i.e. elements with characteristics which are border-line

between the two classes), the so-called *support vectors*; new instances are then classified according to the side of the hyperplane they fall into. In order to handle linearly nonseparable data, the optimization *cost function* includes an error minimization term to penalize the wrongly classified training instances.

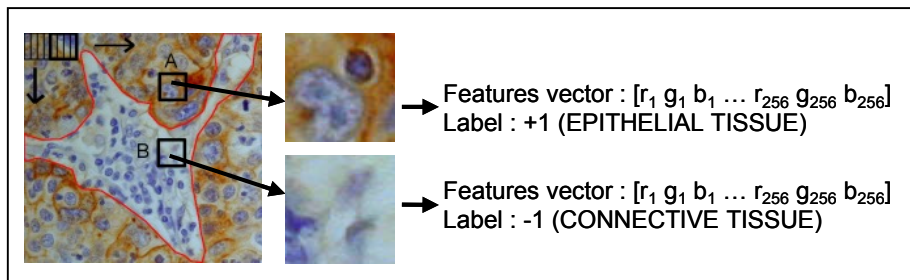
See the references provided in the text for a technical description of SVMs.

Our proposed supervised procedure for tissue segmentation consists in three main steps (see Fig. 5):



**Fig. 5.** Flow-chart of the supervised procedure based on SVM.

**Step 1: Training Features Extraction.** In order to obtain a good generalization of the SVM, a skilled operator was asked to select from a large number of real-life tissue images small rectangular regions wherein both connective and epithelial tissue were present. The images showed various staining levels and very different characteristics of tissue shape and intensity distribution. In each representative sample the operator manually traced the boundaries of connective and epithelial tissue. Then a  $N \times N$  square sliding window was horizontally and vertically shifted over the samples (shift value  $s$ ), thus covering the entire surface of the image; for each shifted window, a features vector was generated with the RGB values of 256 equally-spaced pixels (see Fig. 6, parameters set-up in Section 3.2). In this way, a features vector of  $3 \times 256$  variables was created for each single shift. A +1 label was assigned to windows with a prevalence of epithelial tissue pixels, a -1 label to windows with a prevalence of connective tissue pixels.



**Fig. 6.** Generation of the features vectors for SVM training. A  $N \times N$  square window is horizontally and vertically shifted on the sample, thus covering the entire surface of the image. For each shift a features vector is generated with RGB values of 256 equally spaced pixels, as for *Window A* and *Window B* in the figure above. Epithelial instances are labelled with a +1, connective instances with a -1.



**Step 2: Training.** The labelled features vectors were fed into the SVM for the training; for details about the parameters set-up see Section 3.2.

**Step 3: Classification.** The optimized SVM obtained in the training step is used to perform tissue classification for new images. For this purpose, the input images are processed to generate features vectors as in step 1; then input features vectors are fed into the trained SVM. At the end of the classification, the SVM automatically associates positive labels to epithelial patterns and negative labels to connective patterns. The output is then processed to reconstruct two-dimensional results as those shown in Fig. 8.

### 3 Implementation

The algorithm was implemented in Java as a plugin for ImageJ [20], a public domain image analysis and processing software which runs on all the standard operating systems (Windows, Mac OS, Mac OS X and Linux): therefore it is totally hardware-independent, flexible and upgradeable. We inherited the whole class hierarchy of the open-source ImageJ 1.37 API and the free plugins for color deconvolution [17] and K-means clustering [21] and we implemented our own functions and classes. A user-friendly interface enables the user to set different parameters values without modifying the source code.

For the supervised procedure we used the cSVM tool for binary classification [22], since it uses the state-of-art optimization method SMO, i.e. Sequential Minimal Optimization [23]. This cSVM tool implements the algorithm described in [24], which was successfully used to solve different real world problems. Our ImageJ plugins for features vectors generation and output reconstruction were integrated to the SVM tool.

The parameters of the proposed algorithms were empirically tuned by a skilled operator after running several experiments on a large dataset of real tissue images which showed very different characteristics of staining intensity, resolution, EGF-R/TGF-alpha activity level, tissue shape. In the following subsections, we report some details about the implementation of both the unsupervised and the supervised classification procedures and we outline the experimental set-up of the main parameters.

#### 3.1 Unsupervised Procedure

The *number of clusters*  $k$  (see Section 2.1 step 3) was set to 4 after running the algorithm with values varying from 2 to 5 and evaluating each time K-means performance in terms of sensibility (power to detect connective components) and selectivity (power to avoid misclassification of epithelial components). For values lower than 4 we often experienced a very good sensibility but a not sufficient selectivity; for higher values the sensibility was frequently poor. A  $k$  value equal to 4 assured a good performance of K-means in all the tested images.

The *size threshold*  $T_s$  (see Section 2.1 step 4) was varied from 1000 to 5000 pixels with a step of 1000 and was finally set to 3000. Increasing values led to a progressive improvement of selectivity in the connective tissue selection; with values higher than

3000 the lack in sensibility was often not acceptable. Similarly, the *circularity threshold*  $T_C$  (see Section 2.1 step 4) was decreased from 0,9 to 0,3. A value of 0,7 assured a good selectivity enhancement without altering sensibility in any of the images.

The *edge threshold*  $T_E$  (see Section 2.1 step 5) was increased from 20% to 35% with a step of 5%, evaluating each time the parameter performance in terms of selectivity enhancement and sensibility preservation. A value of 25% assured the best improvement in selectivity without altering sensibility in any of the tested images.

### 3.2 Supervised Procedure

The *window size*  $N$  for features vectors generation (see Section 2.2 step 1) should grant a visible differentiation between connective and epithelial tissue; since nuclei are blue-colored and quite similar in both the tissues, the window has to be large enough to contain a whole nucleus and some surrounding tissue. On the other hand, lower-sized windows allows a better selectivity.

After running several experiments with values varying from 16 to 72 pixels,  $N$  was set to 32 for x200 images and to 64 for x400 images.

Since the optimal window size depends on image resolution, x200 and x400 images were respectively classified with SVM trained with x200 and x400 samples.

The *shift* value  $s$  (see Section 2.2 step 1) was set to  $N/4$ , which granted the best compromise between selectivity of classification and computational time.

After running experiments with linear, gaussian and polynomial kernels, we finally chose the *normalized polynomial kernel* shown in Equation 3, where  $x_1$  and  $x_2$  are feature vectors,  $n=768$  is the input space dimension and  $p=2$  is the kernel hyperparameter; see [24] for technical details.

$$K(x_1, x_2) = \frac{(x_1 \cdot x_2 + n)^p}{\sqrt{(x_1 \cdot x_1 + n)^p} \sqrt{(x_2 \cdot x_2 + n)^p}}. \quad (3)$$

## 4 Experimental Results

We tested the performance of both the algorithms on a large dataset extracted from real tissue images which presented positive reactions at the EGF-R or at the TGF-alpha receptor activation (see Fig. 1 for examples); reactions are localized in cellular membranes for EGF-R and in cytoplasm for TGF-alpha. Images were acquired from different samples with two different enlargements, x200 or x400.

A skilled operator was asked to manually draw the boundaries of connective tissue in each of the testing datasets. The manual segmentations performed by the operator were pixel-by-pixel compared to those obtained by both the unsupervised and the supervised algorithms.

Connective tissue selection was evaluated in terms of *sensibility* (i.e. power to detect connective tissue) and *selectivity* (power to avoid misclassification of non-

connective tissue): for this purpose, the percentage of respectively connective and non-connective pixels which were equally classified by manual and automated segmentation was calculated.

The segmentation *accuracy* was then calculated as weighted average of sensibility and selectivity, as shown in Equation 4.

$$Accuracy = \frac{2}{3} \cdot Sensibility + \frac{1}{3} \cdot Selectivity. \quad (4)$$

Different weights were used because sensibility is more critical for automated measures of protein activity, which is the principal application targeted by our method: in fact, in order to obtain a reliable measure, it is fundamental to eliminate as much as possible non representative tissues from the range of interest; on the contrary, erroneous removal of some epithelial regions is more tolerable, since it has a lower influence on the final measure.

Results obtained for both the automated algorithms are reported in Table 1.

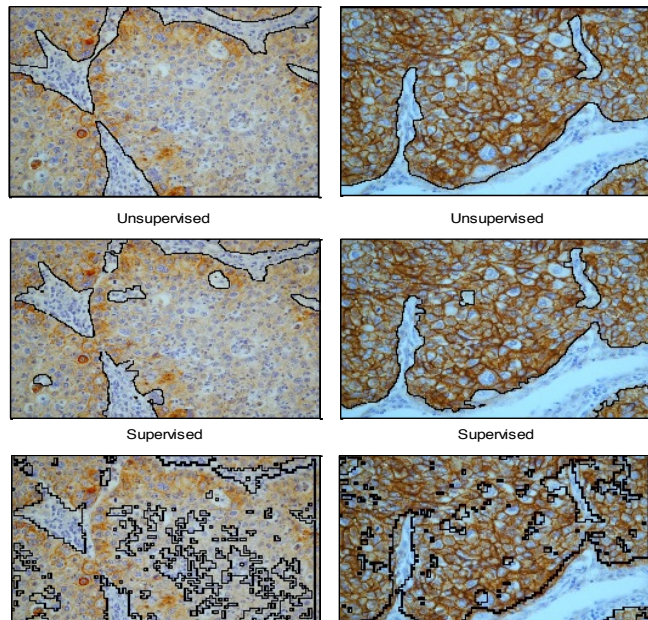
**Table 1.** Experimental results of unsupervised and supervised classifications. As outlined in Section 3.2, in supervised classification two different SVMs trained respectively with x200 and x400 samples were used (the number of training and validation instances extracted from each dataset is reported for both x200 and x400 classifiers). Training instances were removed from the validation dataset, which was considerably larger.

		UNSUPERVISED ALGORITHM			SUPERVISED ALGORITHM				
Dataset		Sensibility (%)	Selectivity (%)	Accuracy (%)	Number of training instances	Number of validation instances	Sensibility (%)	Selectivity (%)	Accuracy (%)
x200	1	81,89	90,54	84,77	1692	28308	57,91	91,38	69,07
	2	94,64	84,94	91,41	912	20263	94,05	79,20	89,10
	3	95,21	97,99	96,14	220	20192	91,09	94,75	92,31
	4	86,60	87,32	86,84	408	19142	84,41	91,18	86,66
x400	5	91,77	86,20	89,91	558	6942	67,48	82,35	72,43
	6	91,30	78,56	87,05	640	6860	66,48	90,02	74,32
	7	99,67	93,33	97,56	252	7248	93,53	87,46	91,51
	8	89,21	86,28	88,23	300	5888	87,29	85,39	86,66

The number of training instances extracted from each dataset is reported too for both x200 and x400 SVMs. The classification performance was evaluated on a large validation dataset (91137 and 28688 validation instances for x200 and x400 SVM, respectively) which did not include the patterns used for training. Some examples of tissue segmentation are shown in Fig 8. As shown in Table 1 and Fig 7, our

unsupervised segmentation process achieved a mean accuracy of 90% in all cases.

**Fig. 7** supervised process



tissue in all 90% mean revised

**Fig. 8.** Examples of tissue segmentation (*in black*). First row: manual segmentation performed by a skilled operator; second row: segmentation performed by the unsupervised procedure; third row: segmentation performed by the supervised SVM approach.

As we previously outlined, SVM is a theoretically superior machine learning method which has often been shown to achieve great classification performance compared to other learning algorithms across most application fields and tasks including image processing [8], [9], [10], [11], [12]. However, in this case its classification performance was poor because of the intrinsic complexity of the images targeted by our method: in fact, these images showed very different characteristics of staining, tissue shape and intensity distribution. Because of the heterogeneity of the representative features of each class, it was impossible for the supervised method to obtain a satisfying separability of connective and epithelial tissue. Images' heterogeneity was less critical for the unsupervised approach, since differently from SVMs it is based only on the characteristics of the input image and not on a fixed model of the ground truth. On the other hand, our unsupervised method's selectivity is influenced by tissue composition: in fact, since the number of clusters is a-priori fixed, some epithelial regions with low brown staining are often misclassified in images without any connective tissue. Despite this eventuality is unlikely, since pure-epithelial tissue samples are very uncommon (and we reasonably suppose that the operator would escape the automated tissue segmentation in this case), we are working on the solution of the problem: in particular, the introduction of an adaptive number of clusters is in development. As regards the supervised approach, other learning methods such as neural networks and artificial neural networks (ANN) will be tested in the future.

## 5 Conclusions

We presented a fully-automated unsupervised tissue image segmentation method that allows to distinguish tumor areas in immunohistochemical images and disregard non pathological areas such as connective tissue. This procedure is critical for automated protein activity quantification in tumor tissues in order to analyze the pathology dynamics and development. We described the original processing steps we designed. Finally, we carried out an extensive experimental evaluation on a large set of heterogeneous images that demonstrated the high accuracy achievable by the proposed technique (90% on average) compared to a more traditional approach based on Support Vector Machines (SVM). As future work, we will compare the proposed approach to artificial neural networks (ANN), and we will eventually study the possibility of their integration.

**Acknowledgments.** We acknowledge the Dep. of Pathology of the S.Luigi Hospital of Orbassano in Turin, Italy, for providing IHC images and for the helpful and stimulating discussions.

## References

1. Taneja, T.K., Sharma, S.K.: Markers of Small Cell Lung Cancer. *World Journal of Surgical Oncology*, vol. 2:10 (2004)
2. Demandolx, D., Davoust, J.: Multiparameter Image Cytometry: from Confocal Micrographs to Subcellular Fluorograms. *Bioimaging*, vol. 5:3, pp. 159--169 (1997)
3. Nedzved A., Ablameyko, S., Pitas, I.: Morphological Segmentation of Histology Cell Images. In: 15<sup>th</sup> International Conference on Pattern Recognition (ICPR00), vol. 1, p. 1500 (2000)
4. Malpica N., de Solorzano, C.O., Vaquero, J.J., Santos, A., Vallcorba, I., Garcia-Sagredo, J.M., del Pozo, F.: Applying Watershed Algorithms to the Segmentation of Clustered Nuclei. *Cytometry*, vol. 28(4), pp. 289--297 (1997)
5. Dybowski, R.: Neural Computation in Medicine: Perspectives and Prospects. *Proc. of the ANNIMAB-1 Conference (Artificial Neural Networks in Medicine and Biology)*, pp. 26--36 (2000)
6. Nattkemper, T.W.: Automatic Segmentation of Digital Micrographs: A Survey. *Medinfo 11(Pt 2)*, pp. 847--851 (2004)
7. Vapnik, V.: *Statistical Learning Theory*. Wiley-Interscience, New York, NY, USA (1998).
8. Angelini, E., Campanini, R., Iampieri, E., Lanconelli, N., Masotti, M., Roffilli, M.: Testing the Performances of Different Image Representation for Mass Classification in Digital Mammograms. *Int. J. Mod. Phys.* 17(1), pp. 113--131 (2006).
9. Osuna, E., Freund, R., Girosi, F.: Training Support Vector Machines: an Application to Face Detection. In: *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'97)*, pp. 130 (1997)
10. Twellmann, T., Nattkemper, T.W., Schubert, W., Ritter, H.: Cell Detection in Micrographs of Tissue Sections Using Support Vector Machines. In: *Proc. of the ICANN: Workshop on Kernel & Subspace Methods for Computer Vision*, Vienna, Austria, pp. 79--88 (2001)
11. Muller, K.R., Mika, S., Ratsch, G., Tsuda, K., Scholkopf, B.: An Introduction to Kernel-Based Learning Algorithms. *IEEE Trans. Neural Networks*, vol. 12:2, pp. 181--201 (2001)
12. Cai, C.Z., Wang, W.L., Chen, W.Z.: Support Vector Machine Classification of Physical and Biological Datasets. *Int.J.Mod.Phys*, vol. 14:5, pp. 575--585 (2003)
13. Ficarra, E., Macii, E., De Micheli, G.: Computer-aided Evaluation of Protein Expression in Pathological Tissue Images. In: *Proc. of IEEE Symposium on Computer-Based Medical Systems (CBMS)*, pp.413-418 (2006)
14. Ruifrok, A.C., Johnston, D.A., 2001.: Quantification of Histochemical Staining by Color Deconvolution. In: *Anal.Quant.Cytol.Histol.*, vol. 23:4, pp.291--299 (2001)
15. Ruifrok, A.C., Katz, R., Johnston, D.: Comparison of Quantification of Histochemical Staining by Hue-Saturation-Intensity (HSI) Transformation and Color Deconvolution. In: *Appl. Immunohisto. M. M.*, vol. 11:1, pp.85--91 (2004)
16. Brey, E.M., Lalani, Z., Hohnston, C., Wong, M., McIntire, L.V., Duke, P.J., Patrick, C.W.: Automated Selection of DAB-labeled Tissue for Immunohistochemical Quantification. In: *J. Histochem. Cytochem.*, vol. 51:5, pp.575--584 (2003)
17. Landini, G.: Software, <http://www.dentistry.bham.ac.uk/landinig/software/software.html>
18. Jain, A.K., Dubes, R.C.: *Algorithms for clustering data*, Prentice Hall (1988)

19. Statnikov, A., Aliferis, C.F., Tsamardinos, I., Hardin, D., Levy, S.: A Comprehensive Evaluation of Multicategory Classification Methods for Microarray Gene Expression Cancer Diagnosis. *Bioinformatics*, vol. 21:5, pp.631--643 (2005)
20. Rasband, W.S.: ImageJ. U.S. National Institutes of Health, Bethesda, Maryland, USA, <http://rsb.info.nih.gov/ij/>
21. Sacha, J.: K-means clustering, <http://ij-plugins.sourceforge.net/plugins/clustering/>
22. Anguita, D., Boni, A., Ridella, S., Riviaccio F., Sterpi, D.: Theoretical and Practical Model Selection Methods for Support Vector Classifiers. Springer. *Support Vector Machines: Theory and Application, Studies in Fuzziness and Soft Computing*, Springer, vol. 177, pp. 159--179 (2005)
23. Platt, J.: Fast Training of Support Vector Machines Using Sequential Minimal Optimization. *Advances in Kernel Methods - Support Vector Learning*. MIT Press, Cambridge, MA, USA (1999)
24. Wang, L.: *Support Vector Machines: Theory and Applications*. Springer, Berlin (2005)