2003 International Symposium on Advanced Intelligent Systems
September 25-28, 2003, Jeju, Korea

Data Mining for Detection of Diabetic Retinopathy

Samuel E. Moskowitz
The Hebrew University of Jerusalem
P. O. Box 7843, Jerusalem 91078, Israel
email:mosk@cc.huji.ac.il

Abstract - **The incidence of blindness resulting from diabetic retinopathy has significantly increased despite the intervention of insulin to control diabetes mellitus. Early signs are microaneurysms, exudates, intraretinal hemorrhages, cotton wool patches, microvascular abnormalities, and venous beading. Advanced stages include neovascularization, fibrous formations, preretinal and vitreous microhemorrhages, and retinal detachment. Microaneurysm count is important because it is an indicator of retinopathy progression. The purpose of this paper is to apply data mining to detect diabetic retinopathy patterns in routine fundus fluorescein angiography. Early symptoms are of principal interest and therefore the emphasis is on detecting microaneurysms rather than vessel tortuosity. The analysis does not involve image-recognition algorithms. Instead, mathematical filtering isolates microaneurysms, microhemorrhages, and exudates as objects of disconnected sets. A neural network is trained on their distribution to return fractal dimension. Hausdorff and box counting dimensions grade progression of the disease. The field is acquired on fluorescein angiography with resolution superior to color ophthalmoscopy, or on patterns produced by physical or mathematical simulations that model viscous fingering of water with additives percolated through porous media. A mathematical filter and neural network perform the screening process thereby eliminating the time consuming operation of determining fractal set dimension in every case.**

I. INTRODUCTION

In terms of human suffering and healthcare costs, diabetic retinopathy is one of the most important ocular consequences of diabetes. Early signs are microaneurysms, exudates, intraretinal hemorrhages, cotton wool patches, microvascular abnormalities, and venous beading. Advanced stages are characterized as proliferative retinopathy. These symptoms consist of neovascularization, fibrous formations, preretinal and vitreous microhemorrhages, and retinal detachment [1].

Microaneurysm count is an indicator of retinopathy progression [2,3,4]. The deeper microaneurysms are detected on fluorescein angiography rather than ophthalmoscopy because of intraretinal edema obscuration [1]. Ophthalmoscopy can take advantage of color differentiation, but the resolution is not adequate to discern pathological lesions of micrometer dimension that appear in early stages of diabetic retinopathy. Another measure is the number of cotton wool patches. Their presence implies further risk to proliferative retinopathy [5]. Fluorescein angiography is therefore the means of choice for *early* detection of microaneurysms and exudates. Being invasive, it is used for selective screening of patients with diabetes mellitus. The purpose of this paper is to apply data mining to detect diabetic retinopathy. Knowledge discovery by machine learning from databases of patterns acquired in routine fundus fluorescein angiography is addressed.

Microaneurysms are intraretinal lesions of linear dimension ranging from 10 μm to 100 μm, spherical or ovoid in shape. They are almost exclusively located on the venous side of the capillary network, and isolated from blood vessels. The latter property distinguishes them from hard exudates or lipid deposits.

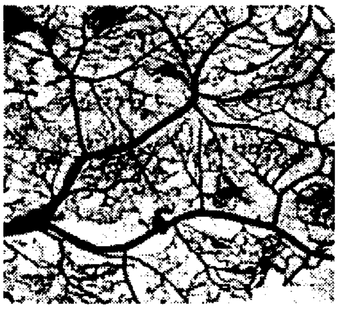

Figure 1. Microaneurysms in diabetes mellitus, adapted from [6].

Complex patterns of blood vessels, microaneurysms, exudates, intraretinal hemorrhages, cotton wool patches, microvascular abnormalities, and venous beading are depicted in Figure 1. Fluorescein angiography shows microaneurysms as hyperfluorescent spots in areas where there is no or little capillary perfusion. They are usually red in color and therefore appear black in red-free photography [7].

Some capillaries become very permeable, so that serum proteins, fluid, and electrolytes can escape, and dye may leak into the retina. Fluorescein angiography can also reveal retinal ischemia related to membrane thinning, nerve fiber loss, and indirect pigment changes as whitened areas.

Images taken on angiography are mathematically filtered to separate microaneurysms. The filter is theoretically based on the distinction between connected and disconnected sets. Discrete sets are given by exudates, microaneurysms, and microhemorrhages. Normal blood vessels, neovascularization and dilated vessels form the connected sets.

II. FILTERING

Filtering isolates microaneurysms, microhemorrhages, and exudates as objects of disconnected sets. The image plane is scanned [8]. Sets of Euclidean space $\mathbf{E} = \mathbf{R}^2$ are examined for connectivity along scan lines [9]. The common definition of connectivity is modified. This concept leads to rules for constructing an algorithm, which are now discussed.

Set M is connected if and only if for every pair of non-void subsets U and V such that $U \cup V = M$

$$(U^{-} \cap V) \cup (U \cap V^{-}) \neq \emptyset \tag{1a}$$

$$|M| = \sup\{|x - y| : x, y \in M\} > 100\ \mu\text{m} \tag{1b}$$

where the super-bar denotes closure. Hence, to be connected it is necessary, but not sufficient, for the set to be connected in the mathematical sense. Sufficiency is reached when the dimension exceeds that of the largest microaneurysm.

If the non-void sets U and V are separated

$$(U^{-} \cap V) \cup (U \cap V^{-}) = \emptyset. \tag{2}$$

Consider sets U and V, both are open or both are closed, then the sets $U - V$ and $V - U$ are separated. The boundary of set U is defined as

$$\text{bd}\,(U) = U^{-} \cap (\mathbf{E} - U)^{-} \tag{3}$$

Several properties are relevant. Suppose U is connected and $U \cap V \neq \emptyset$ as well as $U - V \neq \emptyset$, then $U \cap \text{bd}\,(V) \neq \emptyset$. Points on the boundary are shared. If the set W is connected, and $W \subset U \cup V$, and the sets U and V are separated, then $W \subset U$ or $W \subset V$. Consider connected sets U and V. Then the set $U \cup V$ is connected. Often it is convenient to collect a family of connected and overlapping sets $\{U_\lambda\}$, $\cap_\lambda U_\lambda \neq \emptyset$, then $\cup_\lambda U_\lambda$ is connected. The closure of a connected set is connected, that is, if $U \subset V \subset U^{-}$ and U is connected, then V is connected, as well. If W is a connected subset of connected space $\mathbf{E}$ and $\mathbf{E} - W = U \cup V$, U and V are separated, then the sets $W \cup U$ and $W \cup V$ are connected. The union of two boundary sets in the space of real numbers, all rational numbers and all irrational numbers, is not a boundary set. On the other hand, the union of a boundary set with a closed boundary set is indeed a boundary set.

Another useful definition is the set component of point p

$$C\,(p) = \cup_\lambda U_\lambda \tag{4}$$

where $p \in U_\lambda$ are connected sets, $\forall\lambda$. It is the union of a family of connected sets that contains the same point. Hence, every component is a connected set that is maximal. Moreover, it is closed, for let C be a component, then C^{-} is connected, but C is maximal, therefore $C = C^{-}$. Obviously, two distinct components must be separated. Suppose U is a connected subset of $\mathbf{E}$, and C a component of $\mathbf{E} - U$, then $\mathbf{E} - C$ is connected. The Cartesian product of two connected spaces is connected. What is then concluded for connected sets of $\mathbf{R}$ can be extended to those within the image plane $\mathbf{R}^2$. A set is totally disconnected if $C\,(p) = \{p\}$. This is a mathematical idealization of a cluster containing only isolated microaneurysms.

III. FRACTAL DIMENSION

The progression of the disease is graded by dimension of fractals that remain after filtering. Fractal analysis was used for vascular images [10].

Suppose $\{U_i\}$ is a countable or finite collection of non-empty subsets of $\mathbf{R}^n$ with diameter of most δ that covers F, that is $F \subset \cup_i U_i$, $0 < |U_i| \leq \delta$, then it is a δ-cover of F. Let $F \subset \mathbf{R}^n$ and real number $s \geq 0$, define

$$H^s_\delta(F) = \inf\{\ \Sigma_i |U_i|^s\!:\ \{U_i\} \text{is a } \delta\text{-cover of } F\}. \qquad (5)$$

Hence, for a given δ the smallest summation of different diameters raised to the s-power must be found.

The s-dimensional Hausdorff measure of F is

$$H^s(F) = \lim_{\delta\to 0} H^s_\delta(F). \qquad (6)$$

The limit exists for any subset of $\mathbf{R}^n$.

An outline of the algorithm is as follows: A collection of subsets with $|U_i| \leq \delta$ is selected and fractal F is covered. All diameters are raised to the s-power and then summed. Next, several other arrangements are made and for each placement the procedure is followed again. The smallest summation is $H^s_\delta(F)$. In general, the $\inf_\delta$ depends on the distribution and size of microaneurysms and exudates within the retina. Smaller values of δ are then selected and the entire process repeated. Examining the set of all minima as $\delta \to 0$, $\inf_\delta \to H^s(F)$. The parameter s remains unspecified.

Some useful properties are $H^s(\emptyset) = 0$, $H^s(I) \leq H^s(J)$ for $I \subset J$, and $H^s(\cup_i F_i) = \Sigma_i H^s(F_i)$ provided $\{F_i\}$is a countable collection of disjoint Borel sets. The n-dimensional Hausdorff measure for subsets of $\mathbf{R}^n$ is equivalent to the Lebesgue measure within a scaling factor. Moreover, the Hausdorff measure is scaled according to the rule $H^s(\kappa F) = \kappa^s H^s(F)$ where $\kappa F = \{\kappa x\!: x \in F\}$. Proofs can be found in [11].

It is convenient to envisage this measure in terms of what is needed to determine size. Suppose F is a continuously differentiable 2-dimensional surface in $\mathbf{R}^3$. The 3-dimensional Hausdorff measure is zero because closed balls cannot assess its extent. The 1-dimensional Hausdorff measure is infinite since the figure contains that many lines. For $s = 0$, the measure is also infinite implying the number of points that are required. Although indefinite, the 2-dimensional measure is finite and non-zero.

The Hausdorff dimension of F is defined as

$$\dim_H F = \inf\{\ s\!:\ H^s(F) = 0\} = \sup\{\ s\!:\ H^s(F) = \infty\}. \ (7)$$

At the value of $s^* = \dim_H F$, either $0 < H^s(F) < \infty$, and the Borel set is then an s-set, $H^s(F) = 0$ or $H^s(F) = \infty$.

Suppose $F \subset G$, then $\dim_H F \leq \dim_H G$. Dimension of a mapping $\dim_H f(F) = \dim_H F$, provided f is a bi-Lipschitz transformation, $a\,|x - y| \leq |f(x) - f(y)| \leq b\,|x - y|$ with $x,y \in F$, $0 < a \leq b < \infty$, and $f\!: F \to \mathbf{R}^n$, representing a coordinate translation or rotation in $\mathbf{R}^n$, or an affinity. In addition

$$\dim_H \cup_i F_i = \sup\{\dim_H F_i\}. \qquad (8)$$

Thus, if F is countable or finite, $\dim_H F = 0$.

A set U is closed if $U^- \subset U$ and dense provided $U^- = \mathbf{E}$. If U is the countable set of rational numbers, $\dim_H U = 0$. When its closure is considered, $\dim_H U^- = 1$. The fractal $F \subset \mathbf{R}^n$ and is open, the set has expanse in the entire space therefore the $\dim_H F = n$. In general, for a continuously differentiable m-dimensional manifold, $\dim_H F = m$.

Obviously, the definition is difficult to implement in practice for irregular shapes and therefore other methods are needed. One such approach is called box counting although in this application closed balls are employed. A collection of closed balls is selected with diameter at most δ. Fractal F is covered. The number of balls needed to cover is counted. Other collections whose diameters do not exceed δ are selected. For each collection the number of closed balls is found. Given δ, the smallest number $N_\delta(F)$ is determined. With monotonically decreasing values of δ, more and more irregularities are included as the trial is repeated.

It is possible to graph $\log N_\delta(F)$ versus $\log \delta$. For sufficiently small δ, the plot appears as a straight line provided a power law governs the relationship. Then the slope is equated to

$$\dim_B F = \lim_{\delta\to 0} \log N_\delta(F)/ - \log \delta . \qquad (9)$$

There are lower and upper estimates depending on the nature of covering. If $|F| < \infty$, $F \subset \mathbf{R}^n$, and $N_\delta(F)$ is the smallest number of sets of diameter at most δ that cover F, $\mathrm{ldim}_B F = \mathrm{llim}_{\delta\to 0} \log N_\delta(F)/ - \log \delta$ is the lower and $\mathrm{udim}_B F = \mathrm{ulim}_{\delta\to 0} \log N_\delta(F)/ - \log \delta$ the upper limit, noting that $\mathrm{ldim}_B F \leq \mathrm{udim}_B F$. When $\mathrm{ldim}_B F = \mathrm{udim}_B F$, then $\dim_B F = \lim_{\delta\to 0} \log N_\delta(F)/ - \log \delta$, [11].

IV. NEURAL NETWORK

The neural network is trained on disconnected sets recursively generated from a mathematical or from a physical model of viscous fingering in which water with additives is percolated through porous media. As sources of data, these simulations have distinct advantages. Experimental design can be controlled to reveal pathological specificity and a data warehouse can be filled with unlimited number of images, whereas relatively few archival angiograms exist free of comorbid complexities. The output of the fuzzy computational procedure is fractal Hausdorff and box counting dimensions, a measure of disease progression and therefore grade.

V. CONCLUSIONS

A neural network is trained on the distribution of microaneurysms to return fractal dimension. Hausdorff and box counting dimensions grade progression of the disease. The field is acquired on fluorescein angiography with resolution superior to color ophthalmoscopy, or on patterns produced by physical or mathematical simulations. A mathematical filter and neural network perform the screening process thereby eliminating the time consuming operation of determining fractal set dimension in every case.

REFERENCES

[1] J. Dowler and AM. P. Hamilton, Clinical Features of Diabetic Retinopathy, In: Diabetic Retinopathy, O Paul van Bijsterveld, Editor, Martin Dunitz Ltd, pp. 1-16, 2000.

[2] E. M. Kohner, I. M. Stratton, S. J. Aldington, R. C. Turner, and D. R. Mathews, Microaneurysms in the Development of Diabetic Retinopathy, Diabetologia, vol. 42, pp. 1107-1112, 1999.

[3] D. R. Mathews, E. M. Kohner, S. Aldington, and I. M. Stratton, Relationship of Microaneurysms Count to Progression of Retinopathy Over 6 years in Non-Insulin-Dependent Diabetes, Diabetes, vol. 9, p. 117A, 1995.

[4] R. Klein, S. M. Meuer, S. E. Moss, and B. E. K. Klein, Retinal Microaneurysms Counts and 10-year Progression of Diabetic Retinopathy, Archives of Ophthalmology, vol. 113, pp. 1386-1391, 1995.

[5] Early Treatment Diabetic Retinopathy Study Research Group, Fundus Photographic Risk Factors for Progression of Diabetic Retinopathy, ETDRS Report No. 12, Ophthalmology, vol. 98, pp. 823-33, 1991.

[6] F. W. Newell and J. T. Ernest Ophthalmology Principles and Concepts, The C. V. Mosby Company, Saint Louis, pp. 431-435, 1974.

[7] J. H. Hipwell, F. Strachan, J. A. Olson, K. C. McHardy, P.F. Sharp, and J. V. Forrester, Automated Detection of Microaneurysms in Digital Red-Free Photographs: A Diabetic Retinopathy Screening Tool, Diabetic Medicine, vol. 17, pp. 588-594, 2000.

[8]W. L. Wolfe and G. J. Zissis, Optical-Mechanical Scanning Techniques and Devices, In: The Infrared Handbook, Environmental Research Institute of Michigan, Chapter 10, 1985.

[9] K. Kuratowski, Introduction to Set Theory and Topology, Pergamon Press, Chapter XVII, 1977.

[10] A. Avakian, R. E. Kalina, E. H. Sage, A. H. Rambhia, K. E. Elliott, E. L. Chuang, J. I. Clark, J. N. Hwang, and P. P. Wingerter, Fractal Analysis of Region-Based Vascular Change in the Normal and Non-Proliferative Diabetic Retina, Current Eye Research, vol. 24(4), pp. 274-280, 2002.

[11]K. Falconer, Fractal Geometry Mathematical Foundations and Applications, John Wiley & Sons, New York, pp. 25-52, 1990.