Vol. 29, No. 1/4, 2020

Machine GRAPHICS & VISION

International Journal

Published by
The Institute of Information Technology
Warsaw University of Life Sciences – SGGW
Nowoursynowska 159, 02-776 Warsaw, Poland
in cooperation with
The Association for Image Processing, Poland – TPO

Critical Hypersurfaces and Instability for Reconstruction of Scenes in High Dimensional Projective Spaces

Marina Bertolini[1], Luca Magri[2]
[1]*Dipartimento di Matematica,*
Università degli Studi di Milano, Milano, Italy
[2]*Dipartimento di Elettronica, Informazione e Bioingegneria (DEIB),*
Politecnico di Milano, Milano, Italy
marina.bertolini@unimi.it

Abstract. In the context of multiple view geometry, images of static scenes are modeled as linear projections from a projective space $\mathbb{P}^3$ to a projective plane $\mathbb{P}^2$ and, similarly, videos or images of suitable dynamic or segmented scenes can be modeled as linear projections from $\mathbb{P}^k$ to $\mathbb{P}^h$, with $k > h \geq 2$. In those settings, the projective reconstruction of a scene consists in recovering the position of the projected objects and the projections themselves from their images, after identifying many enough correspondences between the images. A *critical locus* for the reconstruction problem is a configuration of points and of centers of projections, in the ambient space, where the reconstruction of a scene fails. Critical loci turn out to be suitable algebraic varieties. In this paper we investigate those critical loci which are hypersurfaces in high dimension complex projective spaces, and we determine their equations. Moreover, to give evidence of some practical implications of the existence of these critical loci, we perform a simulated experiment to test the instability phenomena for the reconstruction of a scene, near a critical hypersurface.

Key words: critical loci, projective reconstruction, computer vision, multiview geometry.

1. Introduction

As linear projections from $\mathbb{P}^3$ to $\mathbb{P}^2$ are the natural geometric model for images of static three-dimensional scenes captured with pinhole cameras, also linear projections from $\mathbb{P}^k$ to $\mathbb{P}^h$, with $k > h \geq 3$, can be useful in modelling images of particular dynamic and segmented scenes [10, 14, 17, 19, 25, 26, 27]. The classical problem of the *reconstruction* of a static scene – given multiple images of an unknown scene taken from unknown cameras, reconstruct the positions of cameras and of scene points – can be generalized as well in the setting of high dimensional projective spaces. These kinds of problems can be nicely reinterpreted with tools of projective algebraic geometry, which guarantee that sufficiently many images and sufficiently many sets of image correspondences allow for a successful projective reconstruction. The reader is referred to [15] for a wide overview of the role of projective geometry in Computer Vision.

Nevertheless, even in the classical set up of two projections from $\mathbb{P}^3$ to $\mathbb{P}^2$ there are sets of *critical* points, i.e., points for which the projective reconstruction fails, in the sense that for each critical configuration of scene points there exist a non projectively equivalent sets of points and cameras that give the same images in the view planes.

The study of critical loci has been the object of interest for several authors, as shown in literature: in the case of a single view of a *static* scene, where the objective is only the reconstruction of the position of the camera and of the projection matrix (calibration), Buchanan, [8], showed that all the critical configurations lay on a twisted cubic curve. If the scene is static, it is well known that the minimum number of images necessary for a full reconstruction is two. For two views, quadric surfaces were shown to be critical hypersurfaces in [21, 22]. In the case of three or more views, contributions are found in [16,20,23]. A comprehensive, detailed analysis both in the case of two and in the case of multiple views was conducted in [13].

The analysis of *dynamic* or *segmented* scenes has led to the study of projections from higher dimensional space $\mathbb{P}^k$ to the projective plane $\mathbb{P}^2$, as considered by Wolf and Shashua in [27], where the additional dimensions of $\mathbb{P}^k$, with respect to the ambient space, are used to encode information on the evolution of the scene. In this extended space the scene can be treated as static, providing a more manageable representation of dynamic or segmented scenes of the usual space. In this context, critical loci in the case of one view were theoretically described in [6]. The more involved analysis of the critical loci for projective reconstruction from multiple views in higher dimensions is approached in [7] and [2] where the general theoretical framework necessary to describe such critical loci is introduced. This framework showed that critical loci are special algebraic varieties, namely determinantal varieties, and in [3] the authors give a description of the critical loci as zero-sets of suitable ideals. More precisely, revisiting the previous framework in a fully projective context, in [3] critical loci for projective reconstruction in $\mathbb{P}^k$ from n views to $\mathbb{P}^2$ turn out to be either hypersurfaces of degree $\frac{k+1}{2} = n$, if the ambient space is odd dimensional, or special determinantal varieties of codimension 2 and degree $\frac{(k+4)(k+2)}{8}$ if the ambient space is even dimensional, when n is the minimum number of views necessary to allow the reconstruction.

Finally, the notion of criticality can naturally be extended to projections from $\mathbb{P}^k$ to image spaces of higher dimension, $\mathbb{P}^h, h \geq 3$, and the resulting critical loci turn out to be still determinantal varieties whose codimension in $\mathbb{P}^k$ depends on k, h and on the number n of projections.

In this paper the critical locus for n projections from $\mathbb{P}^k$ to $\mathbb{P}^h, h \geq 3$ is studied under the hypothesis that n is the minimal number of projections which allow the reconstruction of the scene, and the dimensions of the ambient space, k, and of the image space, h, are linked by the relation: $k \equiv h - 1 \bmod h$. The interest for this case comes from the fact that, as shown in Section 3, under this numerical hypothesis the critical locus turns out to be a hypersurface in $\mathbb{P}^k$.

This case can be also considered as a generalization of the situation studied in [3] when k is odd. Indeed, in [3] projections are always performed on a projective plane $\mathbb{P}^2$, hence only static images can be modeled. While in this paper we consider projections on spaces of higher dimension $\mathbb{P}^h$ with $h \geq 3$. This generalization allows us to model videos of moving scenes, producing moving images in an image plane which can be treated as a static scene in a projective space of dimension $h \geq 3$.

Even if in our hypothesis the critical locus is a hypersurface hence it has the *higher* dimension allowed in the ambient space, from a practical point of view, it is almost unlikely that all points and all the cameras constitute a critical configuration. Nevertheless, for configuration *close* to critical ones, the attained reconstructions exhibit a certain degree of instability, in the sense that small perturbations of the image points change the reconstructed solution drastically. In order to validate this assertion, following the setup conceived in [7], a simulated experiment for projections $\mathbb{P}^5 \to \mathbb{P}^3$ is performed.

The paper is organized as follows: in Section 2 notations are fixed, some basic definitions from projective algebraic geometry are recalled and a brief introduction to the general computer vision setting is offered for the convenience of the reader. In Section 3 the general theoretical framework for critical configurations and critical loci is described. Section 4 is dedicated to the study of the critical hypersurface: in particular its equation is determined and its singularities are investigated. In Section 5 the instability phenomena are shown in a particular case, i.e. for projections $\mathbb{P}^5 \to \mathbb{P}^3$, with the help of MATLAB [24].

2. General results and preliminaries

In this section we fix notation and terminology, we recall some definitions from projective algebraic geometry which will be useful in the sequel and we give a short overview of classical facts in computer vision related to the problem of projective reconstruction of scenes and cameras from multiple view.

2.1. Notation and basic definitions from Algebraic Geometry

Given a matrix $A = [a_{ij}]$ with real or complex entries, A^T denotes its transpose. The j-th row of A is denoted by $\mathbf{a}^j$. Moreover, $DR^{i_1,\dots,i_n}(A)$ denotes the matrix obtained from A by deleting rows $\mathbf{a}^{i_1}, \dots, \mathbf{a}^{i_n}$.

If $\mathcal{A}$ is the set of the first k integers $\{1, 2, 3, \dots, k\}$, we denote by $\mathcal{A}^{\times n}$ the cartesian product of $\mathcal{A}$ with itself n-times, i.e. $\mathcal{A}^{\times n} = \mathcal{A} \times \cdots \times \mathcal{A} = \{1, 2, \dots, k\} \times \cdots \times \{1, 2, \dots, k\}$.

Now we give some basic definitions in Algebraic Geometry which are useful to understand the following sections. We shall limit ourselves to the case in which the ground field is the field of complex numbers, $\mathbb{C}$. However, for definitions and basic properties concerning projective algebraic varieties, we suggest, for example, [12] or [18].

Following standard notation, $\mathbb{P}^k$ denotes the k-dimensional real (or complex) projective space and $(x_1, x_1, ..., x_{k+1})$ the homogeneous coordinates of its points. Once a projective frame is chosen for $\mathbb{P}^k$, coordinate vectors $\mathbf{X}$ of points in $\mathbb{P}^k$ are written as columns, thus $\mathbf{X}^T = (X_1, X_2, ..., X_{k+1})$. In this context, whenever multiplication by a non-zero scalar is utilized, the scalar will be real or complex, accordingly. A linear projective subspace $\Lambda \subseteq \mathbb{P}^k$ spanned by $m+1$ linearly independent points will be called m-space or subspace of dimension m. By convention the empty set is considered as a (-1)-space.

A *projective algebraic variety* in the projective space $\mathbb{P}^n$ is substantially a subset of points of $\mathbb{P}^n$ defined by the common zeros of a family of homogeneous polynomials.

We need some notions which are basic to study algebraic varieties. Their definitions need a certain amount of technical apparatus, hence we try to give here an informal approach, following [12].

Given a homogenous ideal $I \subset \mathbf{C}[X_1, \ldots, X_{k+1}]$, $V(I)$ denotes the projective algebraic variety defined as $V(I) = \{\mathbf{X} \in \mathbb{P}^k : f(\mathbf{X}) = 0 \text{ for all } f \in I\}$. Details on this standard correspondence between ideals and varieties can be found for example in [9].

An algebraic variety is said to be *irreducible*, if it cannot be expressed as the union of two non-empty proper sub-varieties. Every variety can be expressed as a finite union $\mathcal{X} = \mathcal{X}_1 \cup \mathcal{X}_2 \cup ... \cup \mathcal{X}_r$ of irreducible subsets (subvarieties) of $\mathcal{X}$ which are called irreducible components of $\mathcal{X}$.

The projective varieties contained in $\mathbb{P}^n$ are the closed sets of a topology called the Zariski topology of $\mathbb{P}^n$. The same name will be given to the topology induced by the Zariski topology on the subsets of $\mathbb{P}^n$.

In the Zariski topology, the non empty open sets are very big (they are dense), since the closed sets are the common zeros of some polynomials. This notion is necessary to introduce the notion of *general* (or, sometimes, *generic*). Indeed when a family $\{X_p\}_{p \in \Sigma}$ of objects (points, linear spaces, varieties,...) is parameterized by the points of an irreducible projective algebraic variety Σ, the expression "the general object of $\{X_p\}$ has the property P" means that "the subset of points $p \in \Sigma$ such that the corresponding object X_p has the property P, contains a Zariski open dense subset of Σ" (see for example [12, p. 53]). For example in the family of all the conics in $\mathbb{P}^2$, the general conic is irreducible, since the requirement to be degenerate corresponds to a closed condition in the Zariski topology of the family of conics.

The *dimension*, $\dim(\mathcal{X})$, of an irreducible projective variety $\mathcal{X}$ in $\mathbb{P}^n$ is the integer k such that the general $n-k$-plane of $\mathbb{P}^n$ intersects $\mathcal{X}$ in a finite set of points.

If the variety is reducible, its dimension is the maximum of the dimensions of its irreducible components.

A projective variety defined by a single homogeneous polynomial is called an *hypersurface* of $\mathbb{P}^n$ and it has dimension $n-1$. A hypersurface $\mathcal{X}$ is the zero locus $V(I)$ if I is a *principal* ideal, i.e., an ideal generated by only one element.

The *degree* of a k-dimensional projective variety $\mathcal{X}$ in $\mathbb{P}^n$ is the number (with multiplicity) of points of intersection of $\mathcal{X}$ with a general $(n-k)$-plane of $\mathbb{P}^n$ (from the definition of dimension, this set of points is finite).

For instance a projective curve is a projective variety of dimension one. The degree of a projective curve in $\mathbb{P}^3$ is the number of intersection points of the curve with a generic 2-plane of $\mathbb{P}^3$.

A point $\mathbf{X}$ of an irreducible projective variety $\mathcal{X}$ defined by a family of homogenous polynomial F_i is said to be *singular* if the Jacobian matrix $[\frac{\partial F_i}{\partial x_j}]$ has rank lower than maximum in $\mathbf{X}$. Otherwise the point is a smooth point. The variety $\mathcal{X}$ is smooth if it has no singular points. Notice that in a family of varieties the condition to be singular is a closed condition in the Zariski topology.

When the variety $\mathcal{X}$ is a hypersurface, a point $\mathbf{X} \in \mathcal{X}$ is singular if and only if each line passing through $\mathbf{X}$ intersects $\mathcal{X}$ with multiplicity bigger than 1.

2.2. General setting: scenes, cameras, views

For the convenience of the reader, in this subsections we succinctly recall the concepts of pinhole cameras, centers of projection, views, reconstruction, and critical configurations. For more details we refer the reader to [15] for the classical case of scenes in $\mathbb{P}^3$, and to [2] for the general case of scenes in $\mathbb{P}^k$.

Given a *scene*, i.e., a set of points in the ambient $3D-$space, the action of taking a picture can be modelled by maps that are linear projections from the space of the scene to the plane of the image, the so-called *view*. It is therefore very convenient and natural to assume that the ambient space is embedded in projective $3-$space $\mathbb{P}^3$ and, from the algebraic geometric point of view, it is more convenient to choose a complex ambient space, instead of the real one. Therefore, from now on, all projective spaces are assumed to be complex unless specifically mentioned.

A (pinhole) *camera* can be represented as a central projection P of points in $\mathbb{P}^3$, from a point C, the *center* of the camera, onto the view plane $\mathbb{P}^2$. With respect to the homogeneous coordinates $\mathbf{X} \equiv (X_1, X_2, X_3, X_4)^T$ and $\mathbf{x} \equiv (x_1, x_2, x_3)^T$ in $\mathbb{P}^3$ and $\mathbb{P}^2$ respectively, the projection mapping $P : \mathbb{P}^3 \setminus \{C\} \to \mathbb{P}^2$ can be described by $\mu\mathbf{x} = P\mathbf{X}$, where μ is a non-zero constant and the 3×4-matrix P has maximal rank. The center of projection C is the right annihilator of P. As customary, the projection map and one of its matrix representations in a chosen frame are identified. The set of points in $\mathbb{P}^3$ having the same image under projection P is a line which is called a *ray*.

When several images of the same scene $\{\mathbf{X}_j\}$ are taken with different cameras P_i, $i = 1, \ldots, n$, the images $\mathbf{x}_{ij} = P_i(\mathbf{X}_j), i = 1, \ldots, n$ of the same point via different cameras are called *corresponding* points.

As mentioned above, several authors have introduced generalizations of the classical set up, dealing with certain types of dynamic or segmented scenes, that can be profitably

modelled through the framework of multiple view geometry in higher dimensional spaces. In analogy with the situation in $\mathbb{P}^3$, a *scene* in $\mathbb{P}^k$ is a set of N points $\{\mathbf{X}_j\} \in \mathbb{P}^k$. A *camera* is defined as a projection from $\mathbb{P}^k$ to a projective space $\mathbb{P}^h$ i.e. by a linear map P associated to a full-rank $(k+1) \times (h+1)$ matrix P, whose null space is the *center* of projection. As before, a *ray* is the set of points that are mapped to the same point by P. In this case the center and the rays are linear subspaces of dimension $k-h-1$ and $k-h$ respectively. The notion of corresponding points generalizes to *corresponding* subspaces in the higher dimensional setting: proper linear subspaces $L_i, i = 1 \dots n$, of different views, are said to be *corresponding* if there exists at least a point $\mathbf{X} \in \mathbb{P}^k$ such that $P_i(\mathbf{X}) \in L_i$ for all $i = 1 \dots n$.

2.3. Fundamental matrices and Grassmann tensors

In the classical situation of two cameras P_1 and P_2 taking photographs of a scene $\{\mathbf{X}_j\} \subset \mathbb{P}^3$, the intrinsic relationships between corresponding points in the two view planes are summarized by a 3×3 matrix F of rank 2, the *fundamental matrix* associated to the pair of cameras P_1 and P_2. (see [15] for a thorough exposition.)

Generalizations of the notion of *fundamental matrix* for two view planes in $\mathbb{P}^3$, are given in two different ways. On one side, a *generalized fundamental matrix* is defined in [4] to express the relation between corresponding points in two image spaces $\mathbb{P}^{h_i}, i = 1, 2$, in $\mathbb{P}^k$. On the other side, Hartley and Schaffalitzky in [17], introduced a class of tensors, called *Grassmann tensors*, with the purpose of translating into appropriate equations the relationships among corresponding points, for multiple views in higher ambient spaces. As in the case of the *fundamental matrix*, Grassmann tensors are determined by the projection matrices and, vice versa, the projection matrices can be reconstructed from the Grassmann tensors, up to projective transformation of the ambient space. We recall here the basic elements of their construction and for more details see also [2,17].

Consider a set of projections $P_j : \mathbb{P}^k \setminus C_{P_j} \to \mathbb{P}^{h_j}$, $j = 1, \dots, n$, $h_j \geq 2$ with centers in general position. Moreover consider a *profile*, i.e a partition $(\alpha_1, \alpha_2, \dots, \alpha_n)$ of $k+1$, i.e. $1 \leq \alpha_j \leq h_j$ for all j, and $\sum \alpha_j = k+1$.

Let $\{L_j\}$, $j = 1, \dots, n$, where $L_j \subset \mathbb{P}^{h_j}$, be a set of general s_j-spaces, with $s_j = h_j - \alpha_j$, and let S_j be the maximal rank $(h_j+1) \times (s_j+1)-$matrix whose columns are a basis for L_j. By definition, if all the L_j are corresponding subspaces there exists a point $\mathbf{X} \in \mathbb{P}^k$ such that $P_j\mathbf{X} \in L_j$ for $j = 1, \dots, n$. In other words, there exist n vectors $\mathbf{v}_j \in \mathbb{C}^{s_j+1}$ $j = 1, \dots, n$, such that:

$$\begin{bmatrix} S_1 & 0 & \dots & 0 & P_1 \\ 0 & S_2 & \dots & 0 & P_2 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & \dots & 0 & S_n & P_n \end{bmatrix} \begin{bmatrix} \mathbf{v}_1 \\ \mathbf{v}_2 \\ \vdots \\ \mathbf{v}_n \\ \mathbf{X} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}. \tag{1}$$

The existence of a non trivial solution $\{\mathbf{v}_1, \ldots, \mathbf{v}_n, \mathbf{X}\}$ for system (1) implies that the system matrix has zero determinant. This determinant can be thought of as an n-linear form, i.e. a tensor, in the Plücker coordinates of the spaces L_j. This tensor is the *Grassmann tensor*. In the cases of two views the Grassmann tensor turns out to be the generalized fundamental matrix.

2.4. Projective reconstruction

While reconstruction problems can be posed in several geometric settings as metric, affine, or projective, this work is conducted entirely within the projective framework and therefore reconstruction will always be assumed to be achieved up to projective transformations.

Within a projective setting the camera center is the only property of the camera which is preserved under homographies of the view plane, hence projective reconstruction of cameras consists only of the determination of their centers.

In this subsection we are working under the assumption that the centers C_{P_j} of the projections we consider are in general position. In the examples we will deal with, the technical assumption of centers being in general position implies that $\bigcap_j C_{P_j} = \emptyset$. Notice that reconstruction of a scene would be impossible if $\bigcap_j C_{P_j} \neq \emptyset$ because a scene-point $\mathbf{X}$ would be indistinguishable from any other point in the linear projective space generated by $\mathbf{X}$ and $\bigcap_j C_{P_j}$.

Given n views of a scene $\{\mathbf{X}_j\} \subset \mathbb{P}^k$, the recovery of the scene structure has two consecutive stages: the reconstruction of the camera centers, followed by the reconstruction of the scene, i.e., the position of the points $\{\mathbf{X}_j\}$ in $\mathbb{P}^k$, once cameras have been determined.

To perform both these tasks one needs to have a sufficient number of corresponding points in a suitable number of views. In the classical case of $\mathbb{P}^3$, one easily sees that two views and eight corresponding points allow the reconstruction of the fundamental matrix F by solving a linear system. Once F is determined, projection matrices can also be reconstructed, [15, Section 8.5.3]

In the more general case of multiple views and higher dimensional spaces, reconstruction is significantly more involved and requires the use of Grassmann tensors, see [2,17].

Assuming enough scene points are given, in general enough mutual positions, a first natural question is to determine the minimum number of views necessary to allow reconstruction. In this context two numbers play an important role: the minimum number ω_k of views necessary to reconstruct cameras P_i and a minimum number μ_k of views necessary to reconstruct the scene $\{\mathbf{X}_j\}$ in $\mathbb{P}^k$, when the position of the centers are assumed to be known.

Both these numbers are implicitly given in [17], as the numerical conditions for the

existence of a suitable Grassmann tensor, and explicitly computed in [2] under the hypothesis that all the image spaces have the same dimension h. For the convenience of the reader, we recall them.

Assume that the ambient space is $\mathbb{P}^k$ and all the target spaces have the same dimension, i.e. $h_1 = h_2 = \cdots = h_n = h$. Then the following propositions hold [2]:

Proposition 1. *Assume $k-1 = \sigma h + \lambda$, where σ and λ are non negative integers and $\lambda \leq h-1$. Assuming that cameras are known (up to projective equivalence), the minimum number of views necessary to reconstruct a scene for projections from $\mathbb{P}^k$ to $\mathbb{P}^h$ is*

$$\mu_{k,h} = \sigma + 1\,.$$

Proposition 2. *Assume $k = sh+l$, where s and l are non negative integers and $l \leq h-1$. The minimum number of views necessary to reconstruct the cameras for projections from $\mathbb{P}^k$ to $\mathbb{P}^h$ is*

$$\omega_{k,h} = s + 1\,.$$

2.5. Critical loci

As discussed in the previous section, sufficiently many views and sufficiently many sets of corresponding points in the given views, should allow for a successful projective reconstruction. This is generally true, but it is very easy to notice that even in the classical set up of two projections from $\mathbb{P}^3$ to $\mathbb{P}^2$ one can have non projectively equivalent pairs of sets of scene points and cameras that produce the same images in the view planes, from a projective point of view, thus preventing reconstruction. Such configurations and the loci they describe are referred to as *critical*. Critical loci arising in the reconstruction from a single view, when only the camera can be reconstructed, are fully treated in [6]. A detailed treatment of critical loci in $\mathbb{P}^3$ is found in [13], where the classical result of the criticality of a quadric surface in the case of 2-views, is analyzed.

A partial treatment of critical loci for multiple views in higher dimension is given in [2]. As mentioned in the introduction, in this paper a general framework to study critical loci was proposed, working in a setting in which affine charts had been chosen in each view. Critical loci were shown to be special determinantal varieties, and particular attention was given to the case of $\mathbb{P}^4$ in which a Bordiga surface was obtained as essential component of the critical locus. This case has been further investigated in a fully projective context in [5] and in [1]. Finally, critical loci for multiple views, i.e., for projections from $\mathbb{P}^k$ to $\mathbb{P}^2$, are extensively considered in [3], where the varieties arising as critical loci turns out to be hypersurfaces of degree r in $\mathbb{P}^{2r-1}$ or varieties of codimension 2 and degree $\frac{(r+2)(r+1)}{2}$ in $\mathbb{P}^{2r}$. Moreover, the ideal of these varieties is investigated.

We recall here the formalization of the notion of critical configuration and locus. Let us suppose to have n views of a static scene in $\mathbb{P}^k$, consisting of a set of $N \geq k+3$ points $\{\mathbf{X}_j\}$ in $\mathbb{P}^k$. These n views correspond to n matrices P_i, $i = 1, ..., n$, of dimension

$(h+1)\times(k+1)$ and maximal rank which give the projections $\mathbf{x}_{ij} = P_i(\mathbf{X}_j)$ on the image h-spaces.

Definition 1. *A set of points* $\{\mathbf{X}_j\}$, $j = 1, \dots, N$, $N \geq k+3$, *in* $\mathbb{P}^k$ *is said to be a critical configuration for projective reconstruction from n-views if there exist a non-projectively equivalent set of N points* $\{\mathbf{Y}_j\} \subset \mathbb{P}^k$ *and two collections of* $(h+1)\times(k+1)$ *full-rank projection matrices* P_i *and* Q_i, $i = 1, \dots, n$, *such that, for all i and j,* $P_i\mathbf{X}_j = \mu_{ij}Q_i\mathbf{Y}_j$, $\mu_{ij} \neq 0$. *The two sets* $\{\mathbf{X}_j\}$ *and* $\{\mathbf{Y}_j\}$ *are called* conjugate critical configurations, *with* associated conjugate *matrices* $\{P_i\}$ *and* $\{Q_i\}$.

According to [3], the natural setting to study the locus of all critical configurations associated to sets of conjugate matrices is the product variety $\mathbb{P}^k \times \mathbb{P}^k$, endowed with the two standard projections π_1 and π_2 onto the two factors.

Let $\{\mathbf{X}_j, \mathbf{Y}_j\}$ be conjugate critical configurations as above, with associated conjugate matrices $\{P_i\}$ and $\{Q_i\}$.

Definition 2. *If* $\{(\mathbf{X}_j, \mathbf{Y}_j)\}$ *in* $\mathbb{P}^k \times \mathbb{P}^k$ *are pairs of conjugate critical configurations, with associated conjugate matrices* $\{P_i\}$ *and* $\{Q_i\}$, *the associated* unified critical locus for projective reconstruction from n-views *in* $\mathbb{P}^k \times \mathbb{P}^k$ *is the subscheme* $\mathcal{U}^k = \mathcal{U}^k_{(\{P_i\},\{Q_i\})} \subseteq \mathbb{P}^k \times \mathbb{P}^k$ *defined by the equations* $P_i\mathbf{X}_j = \mu_{ij}Q_i\mathbf{Y}_j$, *given in Definition 1.*

Critical loci appearing in practical applications, and studied in the literature, are the projections of $\mathcal{U}^k$ onto each factor. This motivates the following definition:

Definition 3. *Let* $\mathcal{U}^k$ *be the unified critical locus for projective reconstruction from n-views with associated conjugate matrices* $\{P_i\}$ *and* $\{Q_i\}$, *and let* π_1 *and* π_2 *be the natural projection from* $\mathbb{P}^k \times \mathbb{P}^k$ *onto each factor. The corresponding* critical locus *and, respectively,* conjugate critical locus for projective reconstruction from n-views in $\mathbb{P}^k$ *are the subschemes:*

$$\mathcal{X}^k = \mathcal{X}^k_{(\{P_i\},\{Q_i\})} = \pi_1(\mathcal{U}^k)$$

or respectively

$$\mathcal{Y}^k = \mathcal{Y}^k_{(\{P_i\},\{Q_i\})} = \pi_2(\mathcal{U}^k)\,.$$

3. The critical hypersurface in the case $k \equiv h-1$ mod h

Explicit equations of the critical locus $\mathcal{X}^k$ can be obtained directly making use of the Grassmann tensor introduced in the previous section.

Indeed, the Grassmann tensor $\mathcal{T}^{P_1,\dots,P_n}$ encodes the algebraic relations between corresponding subspaces in the different views of the projections $P_1, \dots, P_n$. Hence by definition of critical set, if $\{\mathbf{X}_j, \mathbf{Y}_j\}$ are conjugate critical configurations, then, for each j, the projections $P_1\mathbf{X}_j, \dots, P_n\mathbf{X}_j$ are corresponding points not only for the projections $P_1, \dots, P_n$, but for the projections $Q_1, \dots, Q_n$, too.

In this section we explicitly construct the Grassmann tensor for n projections from $\mathbb{P}^k$ to $\mathbb{P}^h$, under the hypothesis that $k = h-1$ mod h. Then we use this tensor to

get the generators of the ideal of $\mathcal{X}^k$, which, under our hypothesis, comes out to be a hypersurface.

The condition $k \equiv h-1 \bmod h$, together with the hypothesis that n is the minimum number of views to get a reconstruction, implies that $k = nh-1$ and the only possible *profile* for the Grassmann tensor is $(h, h, \ldots, h)$.

Using the Grassmann formula we get that, if all the centers are in general position, $\dim(\bigcap_j C_{Pj}) = k - n(h+1) < 0$, hence the reconstruction is possible.

In this case $L_1, L_2, \ldots, L_n$ are points and equation (1) specializes to

$$\underbrace{\begin{pmatrix} S_1 & \mathbf{0} & \ldots & \mathbf{0} & P_1 \\ \mathbf{0} & S_2 & \ldots & \mathbf{0} & P_2 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \mathbf{0} & \mathbf{0} & \ldots & S_n & P_n \end{pmatrix}}_{T^{P_1,\ldots,P_n}_{L_1,\ldots,L_n}} \begin{pmatrix} \lambda_1 \\ \lambda_2 \\ \vdots \\ \lambda_n \\ \mathbf{X} \end{pmatrix} = \mathbf{0}\,, \tag{2}$$

where $S_j = (x_{1,j}, \ldots, x_{h+1,j})^T$ are the homogeneous coordinates of points L_j. The left matrix $T^{P_1,\ldots,P_n}_{L_1,\ldots,L_n}$ becomes a square one of dimension $n(h+1) \times n(h+1)$. If in addition $L_1, \ldots, L_n$ are corresponding points, the above linear system has a nontrivial solution $\{\lambda_1, \ldots, \lambda_n, \mathbf{X}\}$ and therefore

$$\det(T^{P_1,\ldots,P_n}_{L_1,\ldots,L_n}) = 0\,. \tag{3}$$

Moreover, the case $\mathbf{X} = \mathbf{0}$ doesn't occur. Otherwise, there would exist a certain $\bar{i}$ for which $\lambda_{\bar{i}} \neq 0$ and we could get $\mathbf{0} = P_{\bar{i}}\mathbf{X} = \lambda_{\bar{i}}\mathbf{x}_{\bar{i}}$ which implies $\mathbf{x}_{\bar{i}} = \mathbf{0}$, a contradiction.

In other words, for the chosen profile $(h, \ldots, h)$, one sees that $\det(T^{P_1,\ldots,P_n}_{L_1,\ldots,L_n}) = 0$ is indeed the n–linear constraint between the homogeneous coordinates $(x_{1,j}, \ldots, x_{h+1,j})$ of the points L_j so to let them be correspondent.

Analogously, if $L'_1, \ldots, L'_n$ is a set of corresponding points in n views, for a set of projections $Q_1, \ldots, Q_n$, we get:

$$\underbrace{\begin{pmatrix} S'_1 & \mathbf{0} & \ldots & \mathbf{0} & Q_1 \\ \mathbf{0} & S'_2 & \ldots & \mathbf{0} & Q_2 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \mathbf{0} & \mathbf{0} & \ldots & S'_n & Q_n \end{pmatrix}}_{T^{Q_1,\ldots,Q_n}_{L'_1,\ldots,L'_n}} \begin{pmatrix} \lambda_1 \\ \lambda_2 \\ \vdots \\ \lambda_n \\ \mathbf{X} \end{pmatrix} = \mathbf{0}\,, \tag{4}$$

and the n linear relation between $L'_1, \ldots, L'_n$ is given by the vanishing of $\det(T^{Q_1,\ldots,Q_n}_{L'_1,\ldots,L'_n})$.

Considering as corresponding spaces $L'_1 = P_1\mathbf{X}, \ldots, L'_n = P_n\mathbf{X}$, expressed in coordinates as $S'_j = (\mathbf{p}^1_j\mathbf{X}, \ldots, \mathbf{p}^{h+1}_j\mathbf{X})^T$, with $\mathbf{X}$ any point in the critical locus, one gets that

the determinant of the following matrix must vanish:

$$M' = \begin{pmatrix} \mathbf{p}_1^1\mathbf{X} & & & & \\ \vdots & \mathbf{0} & \dots & \mathbf{0} & Q_1 \\ \mathbf{p}_1^{h+1}\mathbf{X} & & & & \\ & \mathbf{p}_2^1\mathbf{X} & & & \\ \mathbf{0} & \vdots & \dots & \mathbf{0} & Q_2 \\ & \mathbf{p}_2^{h+1}\mathbf{X} & & & \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ & & & \mathbf{p}_n^1\mathbf{X} & \\ \mathbf{0} & \mathbf{0} & \dots & \vdots & Q_n \\ & & & \mathbf{p}_n^{h+1}\mathbf{X} & \end{pmatrix}. \tag{5}$$

Hence the determinant of M' generates the ideal of the critical locus $\mathcal{X}^k$, as $\mathbf{X}$ has to satisfy no other constraint. So we get that the ideal is principal and we have got the following:

Theorem 1. *Let $k = nh-1$ and let n be the number of views. Then $\mathcal{X}^k$ is a hypersurface of degree n, whose equations is*

$$g = \sum_{(j_1,\dots j_n)\in\mathcal{A}^{\times n}} \mathbf{p}_1^{j_1}\mathbf{X}\cdots\mathbf{p}_n^{j_n}\mathbf{X}\det(DR^{j_1,\dots,j_n}(\mathcal{Q}))\,, \tag{6}$$

where $\mathcal{Q}$ is the $n(h+1)\times(k+1)$ matrix given by staking in column the projection matrices Q_j:

$$\mathcal{Q} = \begin{pmatrix} Q_1 \\ Q_2 \\ \vdots \\ Q_n \end{pmatrix}. \tag{7}$$

As already noted in the Introduction, it is worth observing that the case analysed in Theorem 1 is a generalization to projections to $\mathbb{P}^h, h \geq 3$, of the case, discussed in [3, Section 4] of projections on $\mathbb{P}^2$. Indeed the equation 6 obtained above for the hypersurface $\mathcal{X}^k$ is analogous to the one computed in [3, equation (8)], but the techniques used are very different. Indeed the procedure followed in [3] is much more involved as it conducts to get the generators of the principal ideal via the study of the actions of goups on the maximal minors of suitable matrices. While here we get the generator of the principal ideal via a direct application of the Grassmann tensor.

4. Singularities of the hypersurface $\mathcal{X}^k$

In this section we investigate the singularities of $\mathcal{X}^k$ and we prove the following proposition:

Proposition 3. *The points of $\mathbb{P}^k$ which belongs to at least two center of projections are singular points for the hypersurface $\mathcal{X}^k$, in other words*

$$\bigcup_{i,j=1\dots n} (C_{P_i} \cap C_{P_j}) \subset \mathcal{F}^k \,,$$

where C_{P_i} and C_{P_j} denotes the centers of the projections P_i and P_j, respectively, and $\mathcal{F}^k$ denotes the singular locus of $\mathcal{X}^k$. Moreover if $k \geq 2(h+1)$ then $\mathcal{F}^k \neq \emptyset$, hence $\mathcal{X}^k$ is singular.

Proof. The thesis holds for a generic hypersurface of equation

$$f = \sum_{(j_1,\dots j_n)\in\mathcal{A}^{\times n}} a_{j_1,\dots j_n} \mathbf{p}_1^{j_1}\mathbf{X}\dots\mathbf{p}_n^{j_n}\mathbf{X}\,, \tag{8}$$

where the coefficients $a_{j_1,\dots j_n} \in \mathbb{C}$ are not all zero. Indeed, the structure of the coefficients $a_{j_1,\dots j_n}$, which for the equation of $\mathcal{X}^k$ are the maximal minors of the matrix $\mathcal{Q}$, is not relevant for the implication of the proposition; hence in the following we will consider a hypersurface $V(f)$ for arbitrary coefficients $a_{j_1,\dots j_n}$.

First we can notice that all the projection centers C_{P_i}, $i = 1\dots n$, lies on $V(f)$. Indeed each C_{P_i} is a $(k-h-1)$-linear subspace of $\mathbb{P}^k$, given by

$$C_{P_i} = \bigcap_{j=1\dots h+1} V(\mathbf{p}_i^j\mathbf{X})$$

and, for each fixed i, every summand of f contains one $\mathbf{p}_i^j\mathbf{X}$ as a factor.

Then we show that $C_{P_i} \cap C_{P_j} \subseteq \mathcal{F}^k$, for each $i, j = 1\dots n, i \neq j$. Indeed, $\mathcal{F}^k$ is the set of points $\bar{Y} \in \mathcal{X}^k$ such that each line passing through $\bar{Y}$ intersects $\mathcal{X}^k$ with multiplicity bigger than 1.

Let $l =< \bar{Y}, \bar{Z} >= \{\lambda\bar{Y} + \mu\bar{Z} | (\lambda : \mu) \in \mathbb{P}^1\}$ a line through $\bar{Y}$; the intersection points $l \cap \mathcal{X}^k$ are computed via the solutions λ and μ of the equation:

$$\begin{aligned}
&\sum_{(j_1,\dots j_n)\in\mathcal{A}^{\times n}} a_{j_1,\dots j_n}\mathbf{p}_1^{j_1}(\lambda\bar{Y}+\mu\bar{Z})\dots\mathbf{p}_n^{j_n}(\lambda\bar{Y}+\mu\bar{Z}) = \\
&\sum_{(j_1,\dots j_n)\in\mathcal{A}^{\times n}} a_{j_1,\dots j_n}(\lambda\mathbf{p}_1^{j_1}(\bar{Y})+\mu\mathbf{p}_1^{j_1}(\bar{Z}))\dots(\lambda\mathbf{p}_n^{j_n}(\bar{Y})+\mu\mathbf{p}_n^{j_n}(\bar{Z})) = \\
&\lambda^n \sum_{(j_1,\dots j_n)\in\mathcal{A}^{\times n}} a_{j_1,\dots j_n}\mathbf{p}_1^{j_1}(\bar{Y})\dots\mathbf{p}_n^{j_n}(\bar{Y}) \\
&+\lambda^{n-1}\mu \sum_{(j_1,\dots j_n)\in\mathcal{A}^{\times n}} a_{j_1,\dots j_n}\mathbf{p}_1^{j_1}(\bar{Y})\dots\mathbf{p}_s^{j_s}(\bar{Z})\dots\mathbf{p}_n^{j_n}(\bar{Y}) \\
&+\lambda^{n-2}\mu^2 \sum_{(j_1,\dots j_n)\in\mathcal{A}^{\times n}} a_{j_1,\dots j_n}\mathbf{p}_1^{j_1}(\bar{Y})\dots\mathbf{p}_s^{j_s}(\bar{Z})\dots\mathbf{p}_t^{j_t}(\bar{Z})\dots\mathbf{p}_n^{j_n}(\bar{Y}) \\
&+\dots = 0\,,
\end{aligned} \tag{9}$$

with $(\lambda,\mu)\neq(0,0)$ and $s,t=1,\dots,n$.

Obviously we get that $\bar{Y}\in\mathcal{F}^k \iff \mu=0$ is a double solution of (9) $\iff$ the coefficient of $\lambda^{n-1}\mu$ vanishes for all $\bar{Z}$, being the coefficient of λ^n zero, as $\bar{Y}\in\mathcal{X}^k$. If $\bar{Y}$ belongs to at least two centers, this condition is verified.

Moreover, computing the dimension of $C_{P_i}\cap C_{P_j}$ in $\mathbb{P}^k$ via the Grassmann formula, we get that $\dim(C_{P_i}\cap C_{P_j})\geq 0$ if and only if $k\geq 2(h+1)$, hence, under this assumption, the hypersurface $k\geq 2(h+1)$ is singular. □

5. Experimental validation and instability results

This section is devoted at reporting numerical results to demonstrate the occurrence of instability phenomena near critical loci. Although, from a practical point of view, it is almost unlikely that all points and all the cameras constitute a critical configuration, nevertheless, for configuration close to critical ones, the attained reconstructions exhibit a certain degree of instability, in the sense that small perturbations of the points change the reconstructed solution drastically.

In order to validate the above discussion, following the setup conceived in [7], an experiment for projections $\mathbb{P}^5\to\mathbb{P}^3$ is performed. Specifically, we illustrate the instability of the reconstruction of a dynamic scene modelled by two projections from $\mathbb{P}^5$ to $\mathbb{P}^3$ by performing the following steps.

- *Random generation of projection matrices*
  Two pairs of projections matrices $\{P_1,P_2\},\{Q_1,Q_2\}$ are instantiated: without loss of generality P_1 is chosen as the canonical projection, the remaining projections are randomly generated with integer entries:

$$P_1 = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \end{pmatrix}, \quad P_2 = \begin{pmatrix} -3 & 0 & 0 & -1 & 2 & 1 \\ -1 & -1 & 0 & 1 & 0 & 0 \\ 0 & 3 & -2 & 0 & 0 & -1 \\ 3 & 2 & -2 & 2 & 0 & 0 \end{pmatrix},$$

$$Q_1 = \begin{pmatrix} 0 & -1 & 1 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 1 & -2 \\ 0 & 3 & 0 & 0 & -1 & 0 \\ 0 & 2 & 1 & -2 & -1 & -1 \end{pmatrix}, \quad Q_2 = \begin{pmatrix} 0 & 2 & 0 & 0 & 1 & -4 \\ -1 & 1 & 2 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 \\ 2 & 1 & -1 & 0 & -2 & -1 \end{pmatrix}.$$

- *Equations of the critical locus*

  The ideal of the critical locus for the projection matrices $\{P_1, P_2\}, \{Q_1, Q_2\}$ is determined, using MACAULAY2 [11].

- *Random generation of critical points*

  A set X of 500 points on the critical loci is randomly generated; this set of critical points on the corresponding algebraic set was obtained regarding the defining polynomial as a real valued function and finding randomly distributed zeros through numerical routines in MATLAB [24][1].

- *Perturbation of the critical points*

  The points in X are perturbed with increasing levels of zero-mean gaussian noise, in particular we considered several levels of standard deviation σ_j, and obtained various perturbed configurations. Precisely, we considered 30 different values of standard deviation logarithmically spaced between decades 10^{-16} and 10^{-14}.

- *Projection of the perturbed critical points*

  For each perturbed configuration, i.e. for each $\sigma_j, j = 1, \ldots, 30$, the noisy configurations of points are projected in $\mathbb{P}^3$ using the camera matrices previously introduced.

- *Fundamental matrix estimation*

  Critical points are projected on the two views giving rise to pairs of corresponding points. These correspondence are hence used to estimate a generalized fundamental matrix F_{points}. The procedure [4] employed to compute F_{points} follow closely the classical one to estimate the fundamental matrix in the case of projections from $\mathbb{P}^3$ and $\mathbb{P}^2$: every pair of corresponding points give rise to a constraint on the entries of F_{points} which in turn are determined by solving an overdetermined linear system. As a reference, a generalized fundamental matrix F_{cams} was also computed directly from the cameras. This matrix is not affected by the instability phenomenon.

[1] Code available upon request.

- *Estimating instability*
  Finally we compare the fundamental matrix obtained from correspondences with the one computed from cameras. In order to assess the quality of the reconstructions computed from the perturbed point clouds, we compared the two fundamental matrices, measuring their antipodal distance $d(F_{\text{points}}, F_{\text{cams}})$. In other words, as both F_{points} and F_{cams} are defined up to a multiplicative factor, we identified the space of generalized fundamental matrices with a quotient of the unit sphere in $\mathbb{R}^{16}$ and evaluate the distance between the corresponding two points as:

$$d(A, B) = \min\{\|A - B\|, \|A + B\|\} \tag{10}$$

- *Displaying the results*
  The distribution of these distances with respect to the noise level is reported in Figure 1, where the average angular distance in 1000 trials is reported for each σ_j.

It can be appreciated that when the points of the scene lie near the critical locus – i.e. low values of noise – the instability of the reconstruction ends in the fact that F_{points} is far from F_{cams} and their respective distance are affected by great variance. Therefore, the flawed estimation of F_{points} determines an unreliable reconstruction of a point cloud close to be critical. On the contrary, when the points are far away from the neighborhood

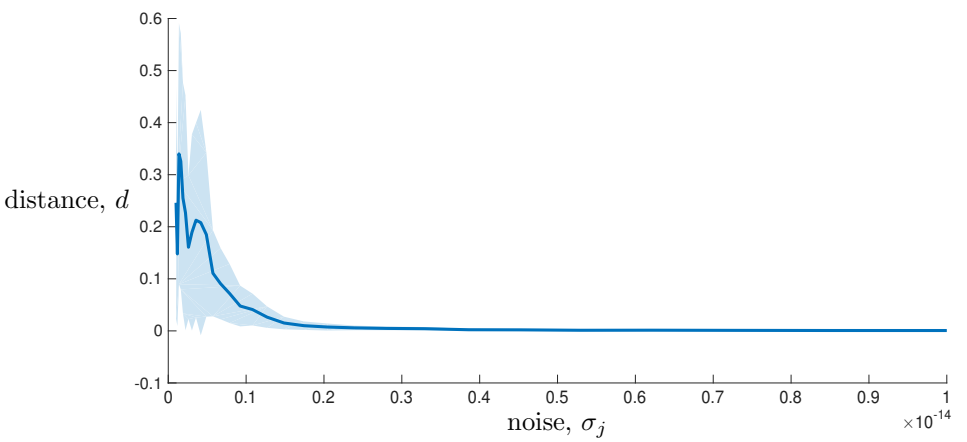


Fig. 1. The antipodal distance $d(F_{\text{points}}, F_{\text{cams}})$, with respect to different levels of noise σ_j, in points on a critical configuration. The average distance is the blue line plot, the width of the shadowed area corresponds to ±standard deviation of the antipodal distances distribution.

of the critical locus – high values of σ – the fundamental matrix F_{points} estimated from the correspondences is consistently close to the reference F_{cams}, and can be profitably used to start the reconstruction process.

This phenomenon is absolutely consistent with the situations analyzed in the other papers [2, 6, 7]: as expected, the larger the distance of points from the critical locus is, the stabler the reconstruction gets.

6. Conclusion

In this paper we study the critical locus for the projective reconstruction of a set of points, in the case of n projections from $\mathbb{P}^k$ to $\mathbb{P}^h$ for $k > h \geq 3$, where n is the minimum number of projections which allows the reconstruction (Propositions 1 and 2) and the dimensions of the ambient space, k, and of the image space, h, are linked by the relation: $k \equiv h-1 \bmod h$. Under this numerical hypothesis (Section 3) the critical locus turns out to be a hypersurface in the ambient space, hence it has the *higher* dimension allowed. The main theoretical result of the paper is contained in Section 3, where, using the notion of Grassmann tensors previously recalled, the equation of the critical hypersurface is obtained in Theorem 1.

Finally, to give evidence of some practical implications of the existence of critical loci, we perform a simulated experiment, in the case of two projections from $\mathbb{P}^5$ to $\mathbb{P}^3$, to show the instability phenomena for the reconstruction of a scene near a critical hypersurface. Indeed, for points close to the critical locus, the attained reconstruction exhibits a certain degree of instability, in the sense that small perturbations of the points change the reconstructed solution drastically.

Acknowledgement

The authors would like to thank Cristina Turrini for helpful conversations.

References

[1] M. Bertolini, G. Besana, R. Notari, and C. Turrini. Critical loci in computer vision and matrices dropping rank in codimension one. *J. of Pure and Applied Algebra*, 224(12):106439, 2020. doi:10.1016/j.jpaa.2020.106439.

[2] M. Bertolini, G. Besana, and C. Turrini. Applications of multiview tensors in higher dimensions. In S. Aja-Fernández, R. de Luis García, D. Tao, and X. Li, editors, *Tensors in image processing and computer vision*, Advances in Pattern Recognition, pages 237–260. Springer, London, 2009. doi:10.1007/978-1-84882-299-3_11.

[3] M. Bertolini, G. Besana, and C. Turrini. Critical loci for projective reconstruction from multiple views in higher dimension: A comprehensive theoretical approach. *Linear Algebra and its Applications*, 469:335–363, 2015. doi:10.1016/j.laa.2014.11.021.

[4] M. Bertolini, G. Besana, and C. Turrini. Generalized fundamental matrices as grassmann tensors. *Annali di Matematica Pura ed Applicata (1923 -)*, Jul 2016. doi:10.1007/s10231-016-0585-4.

[5] M. Bertolini, R. Notari, and C. Turrini. The bordiga surface as critical locus for 3-view reconstructions. *J. of Symbolic Computation*, 91:74 – 97, 2019. MEGA 2017, Effective Methods in Algebraic Geometry, Nice (France), June 12-16, 2017. doi:10.1016/j.jsc.2018.06.014.

[6] M. Bertolini and C. Turrini. Critical configurations for 1-view in projections from $\mathbb{P}^k \to \mathbb{P}^2$. *J. of Mathematical Imaging and Vision*, 27:277–287, 2007. doi:10.1007/s10851-007-0649-6.

[7] M. Bertolini, C. Turrini, and G. Besana. Instability of projective reconstruction of dynamic scenes near critical configurations. In *IEEE Int. Conf. on Computer Vision*, pages 1–7, Los Alamitos, CA, USA, 2007. IEEE Computer Society. doi:10.1109/ICCV.2007.4409100.

[8] T. Buchanan. The twisted cubic and camera calibration. *Computer Vision, Graphics, and Image Processing*, 42(1):130–132, 1988. doi:10.1016/0734-189X(88)90146-6.

[9] D. A. Cox, J. Little, and D. O'Shea. *Ideals, Varieties, and Algorithms: An Introduction to Computational Algebraic Geometry and Commutative Algebra*. Springer, 2015. doi:10.1007/978-3-319-16721-3.

[10] X. Fan and R. Vidal. The space of multibody fundamental matrices: Rank, geometry and projection. In R. Vidal, A. Heyden, and Y. Ma, editors, *Dynamical Vision. Proc. Int. Workshop on Dynamical Vision WDV 2005*, volume 4358 of *Lecture Notes in Computer Science*, pages 1–17. Springer, 2007. doi:10.1007/978-3-540-70932-9_1.

[11] D. R. Grayson and M. Stillman. Macaulay2, *a software system for research in algebraic geometry*. `http://macaulay2.com`.

[12] J. Harris. *Agebraic Geometry: A First Course*, volume 133 of *Graduate Texts in Mathematics*. Springer-Verlag, New York, 1992. doi:10.1007/978-1-4757-2189-8.

[13] R. Hartley and F. Kahl. Critical configurations for projective reconstruction from multiple views. *Int. J. of Computer Vision*, 71(1):5–47, 2007. doi:10.1007/s11263-005-4796-1.

[14] R. Hartley and R. Vidal. The multibody trifocal tensor: motion segmentation from 3 perspective views. In *Proc. 2004 IEEE Computer Society Conf. on Computer Vision and Pattern Recognition CVPR 2004*, volume 1, pages I–I, Washington, DC, USA, 27 Jun-2 Jul 2004. doi:10.1109/CVPR.2004.1315109.

[15] R. Hartley and A. Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, New York, 2nd edition, 2003. doi:10.1017/CBO9780511811685.

[16] R. I. Hartley. Ambiguous configurations for 3-view projective reconstruction. In *Computer Vision. Proc. European Conf. on Computer Vision ECCV 2000, Part I*, volume 1842 of *Lecture Notes in Computer Science*, pages 922–935, Dublin, Ireland, 26 Jun-1 Jul 2000. Springer. doi:10.1007/3-540-45054-8_60.

[17] R. I. Hartley and F. Schaffalitzky. Reconstruction from projections using Grassmann tensors. *Int. J. of Computer Vision*, 83(3):274–293, 2009. doi:10.1007/s11263-009-0225-1.

[18] R. Hartshorne. *Algebraic Geometry*, volume 52 of *Graduate Texts in Mathematics*. Springer, New York, 1977. doi:10.1007/978-1-4757-3849-0.

[19] K. Huang, R. Fossum, and Y. Ma. Generalized rank conditions in multiple view geometry with applications to dynamical scenes. In *Computer Vision. Proc. European Conf. on Computer Vision ECCV 2002, Part II*, volume 2351 of *Lecture Notes in Computer Science*, pages 201–216, Copenhagen, Denmark, 28-31 May 2002. Springer. doi:10.1007/3-540-47967-8_14.

[20] F. Kahl, R. Hartley, and K. Astrom. Critical configurations for n-view projective reconstruction. In

Proc. IEEE Computer Society Conf. on Computer Vision and Pattern Recognition CVPR 2001, pages II–II, Kauai, HI, USA, 8-14 Dec 2001. doi:10.1109/CVPR.2001.990945.

[21] J. Krames. Zur Ermittlung eines Objektes aus zwei Perspektiven (Ein Beitrag zur Theorie der ,,gefährlichen Örter"). *Monatshefte für Mathematik und Physik*, 49:327–354, 1940. doi:10.1007/BF01707311.

[22] S. Maybank. *Theory of Reconstruction from Image Motion*, volume 28 of *Springer Series in Information Sciences*. Springer, Berlin, Heidelberg, 1993. doi:10.1007/978-3-642-77557-4.

[23] A. Shashua and S. J. Maybank. Degenerate n point configurations of three views: Do critical surfaces exist? Technical Report TR 96-19, Hebrew University, 1996. `http://www.cs.huji.ac.il/~shashua/papers/cvpr-critical.ps.gz`.

[24] The MathWorks, Inc. MATLAB. Natick, MA, USA. `https://www.mathworks.com`.

[25] R. Vidal and Y. Ma. A unified algebraic approach to 2-D and 3-D motion segmentation and estimation. *J. of Mathematical Imaging and Vision*, 25(3):403–421, 2006. doi:10.1007/s10851-006-8286-z.

[26] R. Vidal, Y. Ma, S. Soatto, and S. Sastry. Two-view multibody structure from motion. *Int. J. of Computer Vision*, 68(1):7–25, 2006. doi:10.1007/s11263-005-4839-7.

[27] L. Wolf and A. Shashua. On projection matrices $\mathcal{P}^k \to \mathcal{P}^2, k = 3, ..., 6$, and their applications in computer vision. *Int. J. of Computer Vision*, 48(1):53–67, 2002. doi:10.1023/A:1014855311993.

Marina Bertolini is currently Associate Professor of Geometry at the Department of Mathematics at the Università degli Studi di Milano, Italy. Her main field of research is Complex Projective Algebraic Geometry, with particular interest for the classification of projective varieties and for the geometry of Grassmmann varieties. In the last years she has started to work also on applications of Algebraic Geometry to Computer Vision problems, and, in particular, to critical loci for reconstruction and to the geometric properties of multi-view tensors.

Luca Magri graduated in Mathematics at the University of Milan (IT) in 2012. In 2015, he received the PhD from the University of Milan with a thesis on robust multiple model fitting for Computer Vision applications. From 2015 to 2018, he has been a post-doc researcher first at the University of Verona (Dept. of Computer Science) and then at the University of Udine (DPIA), in these periods he collaborated with 3DFlow srl on 3D reconstruction themes. In 2018-2019 he joined the R&D group of FARO Technologies (Rezzato, IT) where he worked on acquisition and registration techniques for structured light scanners. Currently it is at Dipartimento di Elettronica, Informazione e Bioingegneria of the Politecnico di Milano (DEIB) as a postdoctoral researcher.

Text Area Detection in Handwritten Documents Scanned for Further Processing

Jakub Leszek Pach[1], Artur Krupa[2], Izabella Antoniuk[2]
[1]*National Library of Poland, Warsaw, Poland*
j.pach@bn.org.pl
[2]*Institute of Information Technology*
Warsaw University of Life Sciences – SGGW, Warsaw, Poland
artur_krupa@sggw.edu.pl, izabella_antoniuk@sggw.edu.pl

Abstract. In this paper we present an approach to text area detection using binary images, Constrained Run Length Algorithm and other noise reduction methods of removing the artefacts. Text processing includes various activities, most of which are related to preparing input data for further operations in the best possible way, that will not hinder the OCR algorithms. This is especially the case when handwritten manuscripts are considered, and even more so with very old documents. We present our methodology for text area detection problem, which is capable of removing most of irrelevant objects, including elements such as page edges, stains, folds etc. At the same time the presented method can handle multi-column texts or varying line thickness. The generated mask can accurately mark the actual text area, so that the output image can be easily used in further text processing steps.

Key words: text area detection, handwritten text, machine learning, optical character recognition, text recognition.

1. Introduction

Text recognition is a very demanding and varied field of research. Depending on the type of document containing text, i.e., whether it is handwritten or printed, and in the case of handwritten documents, what period is it from, in which region it originated, etc., the methods required to obtain processed text can differ significantly. Furthermore, even before recognizing the actual text, a series of different operations need to be performed, to first optimize the data, remove different types of noise (appearing during data acquisition or occurring in scanned text) and detect the actual text area. Especially in case of handwritten documents these first preprocessing steps are extremely important, since any errors made at this stage can later result in lower accuracy of algorithms used in optical character recognition (OCR).

Processing scanned documents is not a new problem, but its importance is rapidly increasing. When it comes to printed documents, applications such as processing business documents for further use, preparing captioning for hard-hearing persons or voice readings for blind persons come to mind. This problem is even more crucial in the case

of older documents, and untranslated text. Processing handwritten text, especially old or damaged manuscripts, can pose major problems. At the same time, storing the processed text in digital form can speed up its translation, and ease the document circulation between different units (scanned old texts tend to be stored in high resolution while not always containing the amount of information justifying their size). Due to these and other reasons the processing of various documents, especially the handwritten and old ones, have become the objective for researchers and scientists from different fields.

When the text recognition problem for handwritten documents is approached, one of key steps is outlining the area which contains the actual writing. Especially the old documents usually contain a number of irrelevant components, like comments placed on margins, different illustrations, folds, stains, initials, etc., which, if recognized as main text, can actually hinder the quality of OCR made in the following steps. In case of finding the text area, the methodologies used can be divided into three groups: *top-down*, *bottom-up* and *hybrid*. The first approach divides a single image view into smaller parts, to later exclude margins, initials and other elements from the main text. The bottom-up approach groups sets of pixels with a homogenous structure which can be defined using such properties as ink consistency, letter spacing, or similar. Later, single letters are grouped into words, building text lines from them. Hybrid methods use both methodologies, applying the machine learning methods as well as various other computational models to better delimit the text area, and as a result also to improve the text recognition quality.

Among the existing methods, one of the bottom-up procedures divides the binary image (BI) of the processed manuscript by combining the data series encoding the background and ink, to later produce a descriptive rectangle containing only text, without margins or other additional elements [4]. With this procedure both single and double column texts can be processed. In [2] the authors find edges of tables and the margin space. The denoting of the text space is performed by a tracking script to create a curvilinear separation path between each pair of subsequent text lines, which in result leads to finding the separate text fragments. In [8] the text area is identified by first using image binarization and later separating the graph of connected components (CC) with segmentation methods. In a next step, Hough transform is applied to define each connected component and to calculate the distance vector for each graph component, resulting in designating the external edge blocks. Finally, the space found in this way is divided into single text lines by analyzing the CC centroids, which later are grouped into final text space. Another interesting method is applied to Arabic manuscripts analysis [1]. The authors use advanced feature extraction based on image fragments analysis with graph coherent components and a group of multilayer perceptrons, to achieve the highest possible accuracy in separating main text from margins. Finally, a method using Markov random fields (MRF) described in [10] is applied to the analysis of the French manuscript by Gustave Flaubert from XIXth century, to divide the text and the background [13,14].

Taking into account research in the field of finding actual text area in old manuscripts, it can be stated that there still is much space for improvement, especially when it comes to the accuracy line detection, as well as to the speed of the entire process. Our research is inspired by some of these shortcomings and strives to improve the overall text area detection accuracy, while minimizing the number of mistakes at the same time. Results of the presented procedure are meant to be used as an input in the subsequent text processing algorithms.

2. Text area detection

When it comes to text area detection there are many different methods, most of which have some common components. In our case we based our solution on the method described in [17]. The algorithm used in that work is shown in Fig. 1, while each of its key steps is described in subsections below.

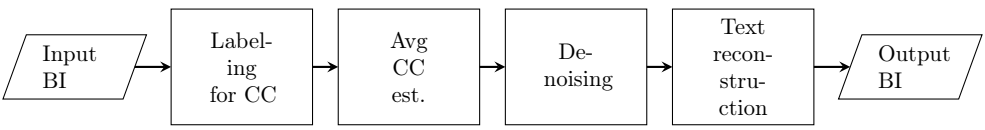


Fig. 1. Stages of the text area detection process.

2.1. Labeling

After we obtain the input binary image (see first stage in Fig. 1), we label each separate CC element in this image, since each of them might be important in the recognition process. In case of text separation we start from finding the text area, which is later divided into lines, words, and, at the final stage, into individual letters. For the examples of connected components please refer to Fig. 2.

There are three ways to describe the CC model (Fig. 3):

- based on the *bounding box* wrapped around the text [7],
- based on the *convex hull* related to the analyzed text [5],
- based on the *ellipse* wrapped around text [6].

The rectangle-based model is more than sufficient for testing purposes as far as performance and efficiency is considered, so this model was used. The binarization was implemented with the classic Otsu method [11]. Labeling the elements of the manuscript requires that the procedure makes it possible to distinguish between three types of CC:

- large elements – blocks of text, stains, folds, shadows, initials;
- medium elements – separated words, letters and fragments;
- small elements – language-specific marks like accents, diacritic elements, dots or unclassified noise.

a

1	1	1	0	1	1	1
1	1	1	0	0	0	1
1	0	1	0	0	0	1
1	0	0	0	1	0	1
1	0	1	1	1	0	0
1	0	1	1	1	0	1
1	0	0	1	0	0	1

b

1	1	1	0	2	2	2
1	1	1	0	0	0	2
1	0	1	0	0	0	2
1	0	0	0	3	0	2
1	0	3	3	3	0	0
1	0	3	3	3	0	4
1	0	0	3	0	4	4

Fig. 2. Example of CC labeling: (**a**) input binary image; (**b**) labeled output image with color labels [9].

Fig. 3. CC labeling: (**a**) bounding box, after [7]; (**b**) convex hull, after [7]; (**c**) ellipse, according to [16].

2.2. The modified Constrained Run Length Algorithm for noise reduction

Binarization of an image is a process in which pixels are assigned only one of two values – 0 or 1. Within the frames of the text recognition process the value 1 (white) will represent the regions where ink is visible and 0 (black) where it is not. In the process of conversion of a text from a color image to a binary image the white color is assigned to text as well as to all irrelevant elements, like image noise, stains, folds and other unnecessary objects. To reduce such noise, we used the Constrained Run Length Algorithm (CRLA) [3, 17] which was modified to better fit our input data.

In the CRLA the Run-Length Encoding (RLE) method is used for information coding. With this method the string in the form "abbcccddddeeeee" can be defined as a counted string which contains the number of occurrences and the occurring character (e.g., ASCII character). In the case of the presented example string, after encoding it with RLE it might look like "1a2b3c4d5e". Shortening of input strings can cause less data to be transferred (here: 10 characters instead of 15). This algorithm was used for the first time in data compression methods [12, 15].

Coming back to noise removal, three important steps need to be performed:

- replacing strings of bits with pairs containing a value and a correlated number of occurrences,
- defining which of the values should be replaced (setting the threshold for replacement),
- reverting the shortened string description to the original form.

The algorithm is run twice: along rows and along columns of the image, giving two images. In each direction, the strings of ones having the length less than the threshold, are replaced with zeros. In passing the image along rows, as the threshold the average width $\bar{w}$ is used, and along the columns the threshold value is the average height $\bar{h}$. The RLE is helpful, as the lengths of the strings are explicitly contained in the code. In this way, small white gaps are removed. Results for an example row of an image are presented in Fig. 4.

The two resulting images – the effect of CRLA filtering along rows and columns – are composed into one output image by performing the AND operation pixel-wise, which is equivalent to pixel-wise arithmetic product.

The effect of the application of the CRLA method to an input image horizontally and vertically is presented in Fig. 5. The application of CRLA to a scan of a manuscript is shown in Fig. 6, where additionally the vertical and horizontal projections of image intensities are shown.

After performing these operations, the output image contains significantly less noise. At the same time, the shapes of underlying pages and small inscriptions in the margin areas are mostly filled with zeros. Furthermore, the bottom area of the document, where some notes written with different handwriting are placed, is visible equally clearly as the main text. Also, at this point only small amount of noise still remains (i.e., remnants of the edges of underlying pages) and resulting images can be further processed.

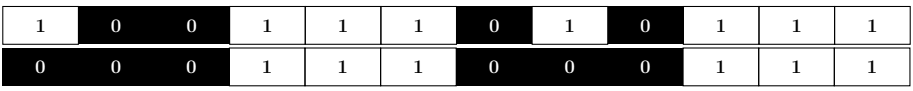


Fig. 4. Input (top) and output (bottom) of CRLA method for an example row.

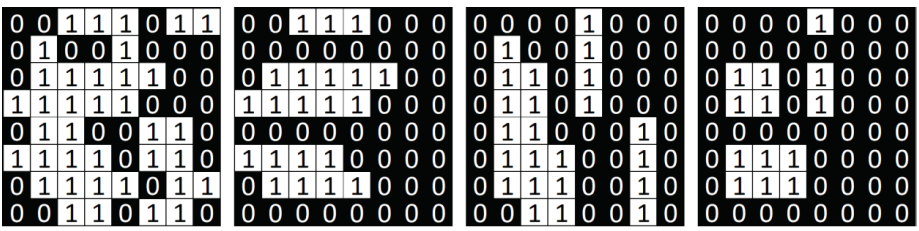


Fig. 5. Results of CRLA filtering: (**a**) input image; (**b**) filtered horizontally; (**c**) filtered vertically; (**d**) output image: pixel-wise product of images **b** and **c**.

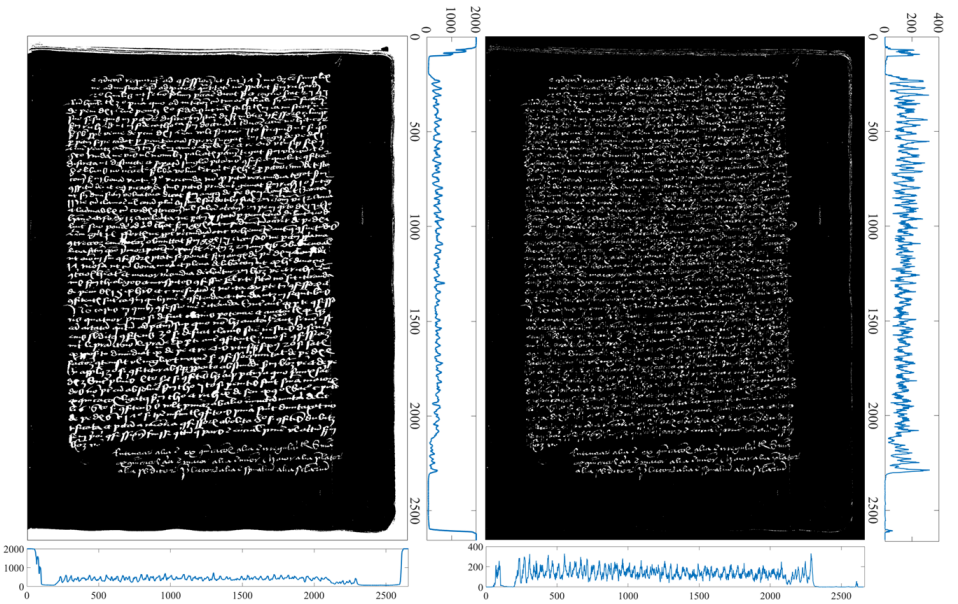


Fig. 6. Noise cancellation in a manuscript: (**a**) before and (**b**) after the CRLA filtering.

2.3. Other noise reduction methods

With the above method, the most part of high-density noise, like stains, folds and similar elements, have been removed. The next step addresses low-density noise elements, as well as unusual objects that can pose some difficulties for the OCR algorithm. Now, the masks of text regions will be generated by classifying image rows and columns as belonging to the text area or not.

To find the threshold for this classification, the average numbers of ones are calculated for each row and column in the binary image. Statistically, when it comes to historical manuscripts, ink would take up to 20% of total page area (in present day documents it would take up to 10%, since nowadays the handwriting is thinner). Therefore, every row (or column) of the image can be safely considered as belonging to text region if it contains more than $\frac{1}{3}$ of white pixels.

The two images, one resulting from classifying the rows, and one from classifying the columns, are combined into the output image by applying the pixel-wise product, as it was done in the previous algorithm. A binary image of a Latin manuscript processed with this method, containing the preliminary mask of text areas, is presented in Fig. 7.

a **b**

Fig. 7. Preliminary text area in a Latin manuscript: (**a**) source and (**b**) result.

3. Image reconstruction

As mentioned before, the input data was a binary image of a manuscript that required preparation for the text recognition process. After performing the above operations, the output is a binary image without noise and without additional objects that could pose problems in further processing steps. Initials, stains, weak ink or punctures present in the original image were filtered out correctly. The final stage of processing is the reconstruction of the handwriting area. This can be compared to typical morphological opening operation. The application of this method to the processing of historical documents was described in [4]. As it was the case with the previous algorithms, as a result we get two images, each being a set of vectors: one with rows and one with columns.

To set the threshold ε for the algorithm, the average height $\bar{h}$ of a connected component in a page will be used:

$$\bar{h} = \frac{1}{n}\sum_{i=1}^{n} h_i \ , \quad \varepsilon = \frac{\bar{h}}{2} \ ,$$

where h_i is the average height of the i-th CC in a page of text, and n is number of all CCs in this page. The rationale for setting the threshold to half of the average height

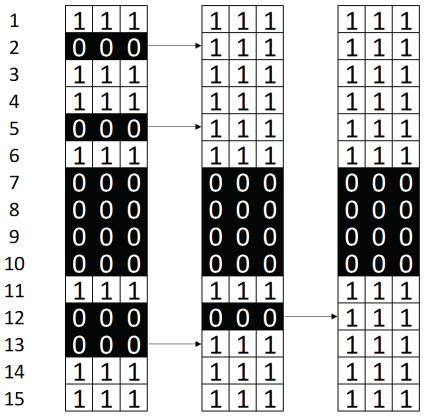


Fig. 8. Three stages of image reconstruction (numbers on the left denote row indexes).

of the connected component is the observation that the height, and also the width, of a typical small letter (like letter 'a', for example) is equal to ε.

The reconstruction goes on according to the following steps, along rows (or columns):

1. Set the `changed` flag to `false`.
2. Set the current element i of the row (or the column) to its first element.
3. It the element i is zero, then count the elements with values one in the neighborhood $[i-\varepsilon, i-1] \cup [i+1, i+\varepsilon]$ of the i-th element (its closed ϵ-neighborhood without the element itself). If their number is greater or equal to ε then change the i-th element to one, and set the `changed` flag to `true`. Go to next element.
4. If not end of row (or column), then proceed from step 3. Otherwise, go to next row (or column).
5. If the row (or column) was the last one, then check the `changed` flag. If `false`, then stop. Otherwise, set the `change` flag to `false`, return to the first row (or column) and proceed from step 2.

The two images, one resulting from performing the above algorithm by rows, and one by columns, are combined into the output image by applying the pixel-wise product, as it was done in the previous algorithms.

Let us consider an example three-column image shown in Fig. 8, where the columns are processed. Assume that $\bar{h} = 4$, so the threshold $\varepsilon = 2$. Let us consider the first column. The first i with a zero is row $i = 2$. Its neighborhood is composed of rows 1, 3 and 4, with three values equal to one; $3 \geq \varepsilon$ so the condition for a change from zero to one is true. The condition is also true for $i = 5$ and 13, so these elements are changed to ones. The same will be done in the remaining two columns. In the next iteration

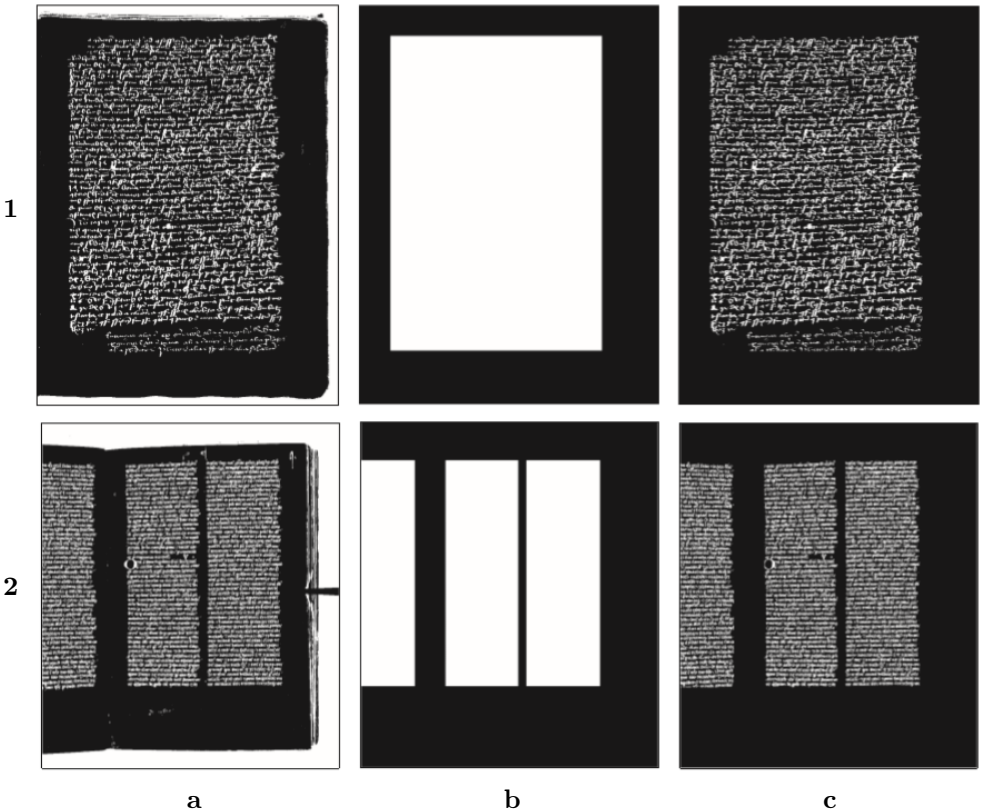


Fig. 9. Intermediate and final results of text area segmentation for two non-trivial manuscripts: (**1**) for a manuscript with differing text thickness and style, and (**2**) for text with multiple columns, uneven spaces and noise produced by page edges and other elements. Stages of processing: (**a**) image after binarization; (**b**) text mask; (**c**) final result after reconstruction.

through this image, the row 12 is changed to one. The image is analyzed one more time and there are no more rows to modify, so the algorithm stops.

Examples of results of the whole text area detection method described in this paper are shown in Fig. 9. Two non-trivial manuscripts are considered: a manuscript with differing text thickness and style, and a manuscript with text in multiple columns, with uneven spaces and noise produced by page edges and other elements. In both cases the unwanted artefacts are properly removed.

4. Conclusion

In this paper we presented a method for text area segmentation for handwritten manuscripts, which can be used as one of the preprocessing steps for further text recognition algorithms. Text recognition is a complex problem, with many difficulties, most of which depend on the type of manuscript and on the final application of the obtained results. The objectives of processing the text and of trying to understand its meaning can vary greatly, from simply storing the data in a most efficient way, up to adjusting the final content to very specific, individual needs.

Our method focuses on the first stage of this process, which is text area detection. Since the algorithms used for text recognition and analysis can be very sensitive to any noise present in the input data, it is crucial to achieve the best possible results at this step. Our method is able to accurately outline the text area, while omitting most of irrelevant elements, such as page edges, stains, folds, etc. At the same time, the presented approach can handle a large level of variety in single manuscripts. Text area can be accurately pointed out in documents with multiple columns, uneven text width, and with different objects not related to actual text. At the same time it also does not cut out the elements such as fragments of text that differ in line thickness. The obtained images are free from most of such elements like margins, spaces between columns, etc., which are irrelevant to the subsequent analysis steps. The obtained final images can be further used in text recognition algorithms.

References

[1] S. S. Bukhari, T. M. Breuel, A. Asi, and J. El-Sana. Layout analysis for Arabic historical document images using machine learning. In *Proc. 2012 Int. Conf. on Frontiers in Handwriting Recognition*, pages 639–644, Bari, Italy, 18-20 Sept. 2012. IEEE. doi:10.1109/ICFHR.2012.227.

[2] M. Bulacu, R. Van Koert, L. Schomaker, and T. van der Zant. Layout analysis of handwritten historical documents for searching the archive of the cabinet of the Dutch queen. In *Proc. 9th Int. Conf. on Document Analysis and Recognition ICDAR 2007*, volume 1, pages 357–361, Parana, Brazil, 23-26 Sept. 2007. IEEE. doi:10.1109/ICDAR.2007.4378732.

[3] B.-S. Chien, B.-S. Jeng, S.-W. Sun, G.-H. Chang, K.-H. Shyu, and C.-H. Shih. Novel block segmentation and processing for Chinese-English document. In *Proc. Visual Communications and Image Processing'91: Image Processing*, volume 1606 of *Proc. SPIE*, pages 588–598, 1 Nov. 1991. doi:10.1117/12.50377.

[4] B. Gatos, G. Louloudis, and N. Stamatopoulos. Segmentation of historical handwritten documents into text zones and text lines. In *Proc. 2014 14th Int. Conf. on Frontiers in Handwriting Recognition*, pages 464–469, Heraklion, Greece, 1-4 Sept. 2014. IEEE. doi:10.1109/ICFHR.2014.84.

[5] S. H. Kim, S. Jeong, G. S. Lee, and C. Y. Suen. Gap metrics for handwritten Korean word segmentation. *Electronics Letters*, 37(14):892–893, 2001. doi:10.1049/el:20010596.

[6] H. I. Koo and N. I. Cho. Text-line extraction in handwritten Chinese documents based on

an energy minimization framework. *IEEE Trans. on Image Processing*, 21(3):1169–1175, 2011. doi:10.1109/TIP.2011.2166972.

[7] G. Louloudis, B. Gatos, I. Pratikakis, and C. Halatsis. Text line and word segmentation of handwritten documents. *Pattern Recognition*, 42(12):3169–3183, 2009. doi:10.1016/j.patcog.2008.12.016.

[8] V. Malleron, V. Eglin, H. Emptoz, S. Dord-Crouslé, and P. Régnier. Text lines and snippets extraction for 19th century handwriting documents layout analysis. In *Proc. 10th Int. Conf. on Document Analysis and Recognition ICDAR 2009*, pages 1001–1005, Barcelona, Spain, 26-29 Jul. 2009. IEEE. doi:10.1109/ICDAR.2009.199.

[9] K. Mirul. Object counting using connected component labelling. In *It's Science – Blog.*, 2020. [Online; accessed 16 Jan. 2020]. `http://k-sience.blogspot.com/2017/06/object-counting-using-connected.html`.

[10] S. Nicolas, T. Paquet, and L. Heutte. Complex handwritten page segmentation using contextual models. In *Proc. 2nd Int. Conf. on Document Image Analysis for Libraries DIAL'06*, pages 46–59, Lyon, France, 27-28 Apr. 2006. IEEE. doi:10.1109/DIAL.2006.8.

[11] M. Otsu. A threshold selection method from gray-level histograms. *IEEE Trans. Systems, Man and Cybernetics*, 9(1):62–66, 1979. doi:10.1109/TSMC.1979.4310076.

[12] J. L. Pach. Analysis of lossless data compression methods (in Polish). Technical report, Warsaw University of Life Sciences – SGGW, Faculty of Applied Informatics and Mathematics – WZIM, Warsaw, 2011.

[13] J. L. Pach. *Identification of the author of Latin manuscripts with the use of image processing methods* (in Polish). PhD thesis, Warsaw University of Technology, Faculty of Electronics and Information Technology, Warsaw, Poland, 2019.

[14] J. L. Pach and P. Bilski. Robust method for the text line detection and splitting of overlapping text in the Latin manuscripts. *Machine Graphics & Vision*, 23(3/4):11–22, 2014. `http://mgv.wzim.sggw.pl/MGV23.html#3-11`.

[15] D. Pountain. Run-length encoding. *Byte*, 12(6):317–319, 1987. `https://archive.org/details/byte-magazine-1987-06`.

[16] J. Ryu, H. I. Koo, and N. I. Cho. Language-independent text-line extraction algorithm for handwritten documents. *IEEE Signal Processing Letters*, 21(9):1115–1119, 2014. doi:10.1109/LSP.2014.2325940.

[17] F. M. Wahl, K. Y. Wong, and R. G. Casey. Block segmentation and text extraction in mixed text/image documents. *Computer Graphics and Image Processing*, 20(4):375–390, 1982. doi:10.1016/0146-664X(82)90059-4.

Skull Stripping Using Traditional and Soft-Computing Approaches for Magnetic Resonance Images: A Semi-Systematic Meta-Analysis

Humera Azam, Humera Tariq
Department of Computer Science, UBIT, University of Karachi, Karachi, Pakistan
humera.azam@uok.edu.pk

Abstract. MRI scanner captures the skull along with the brain and the skull needs to be removed for enhanced reliability and validity of medical diagnostic practices. Skull Stripping from Brain MR Images is significantly a core area in medical applications. It is a complicated task to segment an image for skull stripping manually. It is not only time consuming but expensive as well. An automated skull stripping method with good efficiency and effectiveness is required. Currently, a number of skull stripping methods are used in practice. In this review paper, many soft-computing segmentation techniques have been discussed. The purpose of this research study is to review the existing literature to compare the existing traditional and modern methods used for skull stripping from Brain MR images along with their merits and demerits. The semi-systematic review of existing literature has been carried out using the meta-synthesis approach. Broadly, analyses are bifurcated into traditional and modern, i.e. soft-computing methods proposed, experimented with, or applied in practice for effective skull stripping. Popular databases with desired data of Brain MR Images have also been identified, categorized and discussed. Moreover, CPU and GPU based computer systems and their specifications used by different researchers for skull stripping have also been discussed. In the end, the research gap has been identified along with the proposed lead for future research work.

Key words: skull stripping, brain MR Images, soft computing, meta-analysis.

1. Introduction

The rich advancement in computing world has made it easier for medical experts to diagnose a particular disease or abnormality in living bodies. There are numerous computer aided diagnostic techniques which are helping doctors, bio-scientists and other medical investigators to understand the novel issues and their proposed solution. Image processing is the backbone of any computer aided mechanism and there are numerous techniques being used in medical field to investigate the human body out of which common techniques are X-rays, Computed Tomography, Magneto Encephalography, Positron Emission Tomography, and the most common and popular technique is the Magnetic Resonance Imaging (MRI) [29, 54].

Primary competitive advantages of using MRI over other types include its quality of being non-invasive and the fact that it provides more detailed, deep and comprehensive

images of organs [3] than the majority of other methods. There are four common modalities of MR images, including Longitudinal Relaxation Time (T1), Transverse Relaxation Time (T2), Proton Density (PD) and Fluid Attenuated Inversion Recovery (FLAIR) [62].

Scanners of MRI scan the body and create numerous images from multiple rotated axes, due to which, different views are reported for diagnosis. The 3D nature of MRI helps taking the view of the body from left to right, top to down, and from front to back [3,4,56]. The common types of anatomical orientation are Coronal plane from front to back; Sagittal plane from left to right; and Transversal plane from top to down [56].

The brain is a very sensitive part of the human body as it is made of soft tissues which are a combination of cerebrospinal fluid and fats. Such a complex system is fully covered with the strongest bone of the body called the skull [44]. MRI scanners capture the skull which needs to be removed for clearer understanding of the actual brain tissues [50]. The process of removing the skull from the brain images is called skull stripping. The more precise and efficient skull stripping ensures better help for clinical diagnosis.

This research study consists in the review of existing methods available for skull stripping from brain MR images along with their merits and demerits. Moreover, identifying the research gap in order to understand the current status and get lead for future research work also belongs to the scope of the present study.

1.1. Significance of the study

This research study provides understanding of the existing research gap and provides an abstraction of the experimental framework for future experiments generally in the field of *digital image processing* and most specifically in the domain of *brain MR images* for removing skull and other non-brain cells, in order to enhance the readability and understanding of brain MR images by medical experts for diagnostic purposes.

1.2. Methodology

This research study is carried out using semi-systematic review of literature pertaining to skull stripping methods. Fully systematic review requires extensive resources as well as at least 18 months to complete. Both said constraints provided the rationale to opt for the semi-systematic approach instead of the fully systematic one. Reviewed research studies are available in the respective cited journals for analyses using the meta-synthesis approach. Thematic convergence of different skull stripping methods has been assessed as the outcome of meta-synthesis on the basis of shared properties or architectural similarity between them.

2. Thematic convergence of skull stripping methods

Thematic convergence of developed, reviewed and discussed methods of skull stripping and image processing by different authors in latest research studies has been discussed in temporal order, i.e. older to newer.

2.1. Traditional methods

The reign of traditional methods have been popular in the field of image processing until the invention of neural networks. The convergence of traditional methods has been synthesized in the following subsections.

2.1.1. Traditional methods recently studied

Histogram Analysis and Deformable Model methods comprising the Thresholding and Simplex Mesh respectively have offered significantly positive results on the scale of Jaccard Index = 0.904, Dice Similarity Coefficient (DSC) = 0.95, Specificity = 0.985 [17]. Researchers experimented with Multi Atlas method [11] and Atlas model [20] with significant results of DSC = 0.9802, Specificity = 0.9908, Sensitivity = 0.9802, Average Distance = 0.66 and Hausdorff Distance = 7.72. Binarization method [40] including the irrational filter has provided significant results on the scale of DSC = 0.942, Sensitivity = 0.912, Specificity = 0.971, Overlap Fraction = 0.958 and Extra Fraction = 0.092. The said method remained competitive to the Otsu's method [53]. Another traditional method named as S3 [48] based upon brain anatomy and image intensity has also provided significant results on the scale of Jaccard Similarity $>$ 0.99 and 0.95 for data-sets taken from BrainWeb [5, 6] and IBSR [59] databases respectively; moreover, three measures of DSC, Sensitivity and Specificity $>$ 0.99 for both data-sets. Mathematical Morphology [2] based upon erosion and dilation have also provided better results for skull stripping.

The summary of above discussed traditional methods recently experimented with is presented in Table 1.

2.1.2. Competitive methods in comparison with traditional methods

Common state of the art competitive methods in comparison with traditional methods include Brain Extraction Tool (BET) [11,17,48], Brain Surface Extractor (BSE) [17,48], Robust Brain Extraction (ROBEX) [11,48]. Afore-cited research studies have offered better results in terms of performance measures such as Precision, Accuracy, Effectiveness and Efficiency (PAEE) while comparing with aforementioned state of the art methods.

2.1.3. Data and systems used for traditional methods

Most common data-sets taken for experimenting with the most recently tested traditional methods include Internet Brain Segmentation Repository (IBSR) [2, 11, 17, 40, 48, 59],

Tab. 1. Summary of traditional methods

Author & year	**Methods studied**	**Backbone architecture**	**Measures calculated**	**Methods compared**	**Data type**
Galdames et al. (2012) [17]	Histogram Analyses and Deformable Model	Thresholding and Simplex Mesh	Jaccard Index .904; DSC .9500; Specificity .985; Sensitivity .9900	HWA; BET and BSE	T1 from BrainWeb and IBSR
Doshi et al. (2013) [11]	Multi Atlas Model	Single Atlas and Multi Atlas	DSC .9802; Specificity .9908; Sensitivity .9802; Average Distance .66; Hausdorff Distance 7.72	BET and ROBEX	T1 from ADNI; IBSR and OASIS
Huang and Parra (2015) [20]	Atlas Model	Unified Segmentation Algorithm	Tissue Correlation Map	Intra-method	T1 from BrainWeb and Marom Bikson
Moldovanu et al. (2015) [40]	Binarization Mehtod	Irrational Filter	DSC .942; Senstivity .912; Specificity .971; Overlap Fraction .958; Extra Fraction .092	Otsu [42]; Sauvola [51]; Niblack [41]; Bernsens [1] methods	T1; T2; GAD and PD from WBA; T2 from IBSR
Roy and Maji (2015) [48]	S3	Brain Anatomy and Image Intensity	Jaccard Similarity .99 for BrainWeb and .95 for IBSR; DSC .99; Senstivity .99; Specificity .99	BET; BSE and ROBEX	T1 from BrainWeb; T1 from IBSR
Bhadauria et al. (2020) [2]	Mathematical Morphology	Erosion and Dilation	N/A	Intra-method	WBA and IBSR

BrainWeb [5, 17, 20, 48], and Open Access Series of Imaging Studies (OASIS) [11, 27, 28]. Only T1 weighted brain MR images both simulated and real have been used for the purpose. CPU based systems with 8 GB RAM have been used by the number of researchers for experimenting with traditional methods.

2.2. Deep Learning Neural Network based methods

Deep Learning Neural Network (DLNN) based methods took over the reign of traditional methods because of their enhanced sophistication with their own strengths and weaknesses. The convergence of recently studied DLNN based methods has been synthesized in the following subsections.

2.2.1. Recently developed DLNN methods

Through numerous experiments, the robustness of DLNN based architectures including U-Net, Rectified Linear Unit (ReLU), ConvNet, ResNet, and ConsNet has been proved.

Intensive review has suggested that the most common architectures include U-Net [7, 12, 14, 21, 22, 23, 30, 36, 37, 55].

U-Net architectures of both 2D and 3D types have successfully produced significant results for different performance measures of PAEE in different research studies. In an experimental research study, DSC = 0.71 has been achieved while utilizing the following hyperparameters: Epochs = 4, Discount Rate = 0.5 and 0.2, and Learning Rate = 0.0004 [14]. In another study, researchers have achieved DSC = 0.965 with False Negative Rate (FNR) = 0.2 and False Positive Rate (FPR) = 0.8 by implementing three layers of Convolutional Neural Network (CNN) with one steroid in the first and two steroids in the second layer [55]. Simultaneous Truth and Performance Level Estimation (STAPLE) constituted over 2D FCN U-Net has achieved DSC = 0.9575, 0.8887 and 0.8932 for three different data-sets of T1 weighted MR images with Learning Rate = 0.0001; while the measures of Sensitivity, Specificity, Hausdorff and Mean Distance were also significant [36]. The version of 2D U-Net has been extended for establishing 3D U-Net through max-pooling and batch normalization, which has achieved DSC = 0.9903, Sensitivity = 0.9853 and Specificity = 0.9953 on the data-set of T1 weighted MR images [21]. Researchers have experimented with the method HD-BET which is primarily comprised of U-Net CNN with remarkable results for the measures of DSC = 0.976 and Hausdorff Distance = 3.3 using T1, T2 and FLAIR images from databases of European Organization for Research and Treatment of Cancer (EORTC), LONI Probabilistic Brain Atlas (LPBA) and Neurofeedback Skull-stripped (NFBS) [22]. Researchers experimented with 3D U-Net based method comprised of Transfer Learning (TL) and Multi Output Net which performed exceptionally with DSC = 0.785 and 0.843 on the data-set of Multi-Atlas Labeling Challenge (MALC) and Hammers Adult Atlases (HAA), respectively [7]. Researchers experimented with another 2D U-Net based method of STAPLE which offered high rates of DSC = 0.9718 and Symmetric Surface-to-Surface Mean Distance (SSSMD) = 0.037 on T1 weighted images taken from databases of Calgary-Campinas, LPBA and OASIS [37]. The score of other scales like Sensitivity = 0.9891, Specificity = 0.9946 and Hausdorff Distance = 9.713 have also been remarkable but could not outperform other state of the art methods in comparison. Different hyperparameters have been used for the experiment including Learning Rate = 0.001, Exponential Decay = 0.995 after each epoch, and Fixed Kernel Size = 3×3 [37]. Time Distributed U-Net based CNN method has been tested with Model Accuracy = 0.583 in intra-method comparison with T1 weighted images taken from the database of MICCAI Brain Tumor Segmentation (BraTS) [12]. Researchers experimented with the method of Cascade 3D U-Net based CNN while using hyperparameters of Learning Rate = 10 – 5, Weight Decay = 0.0005, Momentum = 0.9 (in Adam optimizer), and Epochs = 300 [23]. The method offered considerably good results and achieved Root Mean Square (RMS) = 0.86 on 90 MR images of kidney. In another research study, an experiment with the method of U-Net based CNN named as ACEnet has been carried out with hyperparameters like Epochs

= 100, Dropout Rate = 0.1, Momentum = 0.9 and Weight Decay = 0.0001 [30]. The studied method offered remarkable results as DSC $\geq$ 0.8 and Average Time to Segment $\approx$ 10 s on T1 weighted MR images taken from databases of MALC, Alzheimer's Disease Neuro-imaging Initiative (ADNI), Mindboggle, and SchizBull (see [30] for references).

ReLU architectures have also successfully produced significant results for different measures of PAEE in different research studies. An experiment has been run with ReLU architecture and achieved significant results as DSC = 0.965, FNR = 0.2 and FPR = 0.8 using T1 weighted images taken from NFBS [55]. Apart from this, an experiment has been carried out with ReLU based CNN which provided remarkable results for the measure of Sensitivity $>$ 0.87, Specificity $>$ 0.94 and Accuracy $>$ 0.918 on T1 weighted images taken from OASIS [52]. Another ReLU based CNN named as DeepMedic performed outstanding using hyperparameters of Learning Rate = 0.0005 and Epochs = 35 on T1 weighted MR images taken from different data-sets of OASIS, LPBA, and St. Olavs Hospital [13]. ReLU has also been included in an experiment along with U-Net features and achieved significant results [21]. An experiment has been carried out on ReLU based CNN named as DeepICE using hyper-parameter of Epochs = 20 with significant results of DSC = 0.9889 on T1 weighted MR images taken from IXI, OASIS, and BSTP [38]. CNN based methods of Focal Loss and RetinaNet based upon multiple architectures like ReLU, ConvNet, and ResNet have been experimented with using hyperparameters of Learning Rate = 0.01×0.1 after 60 K and then after 80 K iterations, Momentum = 0.9 and Weight Decay = 0.0001 [31]. The method tested increased the mean Average Precision 3-4 points on each T1 weighted MR image taken from Common Objects in Context (COCO) [33].

The summary of the above listed DLNN methods is presented in Table 2.

2.2.2. The rise of masking technique in DLNN methods

Along with the success of U-Net and ReLU based DLNN, another great architecture ResNet jointly with Region CNN R-CNN and in the latest cases with Faster R-CNN methods [45, 46] has provided significant results in numerous experiments. The state of the art method of Mask R-CNN [32] has been tested which is primarily based upon the architecture of Faster R-CNN, Feature Pyramid Network (FPN), ResNet, and ResNeXt, and is using hyperparameters of Learning Rate = 0.02, Weight Decay = 0.0001, and Momentum = 0.9 on T1 weighted MR images taken from COCO [18]. Before this, the FPN has been studied which has later been induced to postulate and experiment the revolutionary method of Mask R-CNN [32]. The developed FPN is based upon Faster R-CNN and two versions of ResNet50 and ResNet101 with hyperparameters of Learning Rate = 0.02×0.1 after 60 K and 80 K iterations on T1 weighted MR images from COCO [33] and PASCAL [15]. In continuation of their own work, researchers experimented with RetinaNet which actually received the contribution from their own FPN [31]. Transfer

Learning in Mask R-CNN has successfully been induced with hyperparameters of Learning Rate = 0.02×0.1 after 60 K and then 80 K iterations on T1 weighted MR images from COCO and Visual Genome [19]. Non-local Neural Network functionally comprising Mask R-CNN and ResNet architectures has been tested with hyperparameters of Learning Rate = 0.01×0.1 after every 150 K iterations, Momentum = 0.9, and Weight Decay = 0.0001 [58]. Apart from the novelty of the method, the experiment is unique because the video data has been taken into experiment for segmenting moving objects.

2.2.3. Competitive methods in comparison with DLNN methods

DLNN methods have outperformed traditional methods [16] out of which prominent DLNN methods include Bayesian Evolutionary Analysis by Sampling Trees BEaST [24, 37, 38, 47, 49], ROBEX [21, 22, 24, 37, 47, 49, 55], BET [22, 24, 37, 49], Hybrid Watershed Algorithm (HWA) [24, 37], BSE [22, 24, 37, 49, 55], FMRIB Software Library (FSL) [55], Analysis of Functional NeuroImages (AFNI) [47, 55], Advanced Normalization Tools (ANTs) [22,55], CompNet [10], Spectre [47], Kleesiek's method [21], 3dSkullStripping [22, 24], SLAN [7], Marker based Watershed Scalper (MBWSS), STAPLE and Optimized Brain Extraction Tool (OptiBET) [37], FreeSurfer [57], NICE [38], G-RMI [31, 32], and AttractioNet [32].

2.2.4. Data and systems used for DLNN methods

Experimental studies conducted to test different DLNN methods of skull stripping has taken data from different databases out of which some are publicly available and for the rest of them the prior permission is needed to access the database and to use data. Leading databases provided different types of brain MR images like T1 weighted, T2 weighted, FLAIR etc. and such databases include OASIS [10, 13, 24, 36, 37, 38, 52], IBSR [24, 57], LPBA [13,22,24], MALC [7,30], ADNI [30], PASCAL [19,32], COCO [18,19,31,32], Hammers [7], NAMIC [49], MPRAGE [49], UKBB [7], BraTS [12, 14], Visual Genome [19], NFBS [21, 22, 55], and Calgary-Campinas, [22, 36].

In addition to databases, different GPU based computer systems have been utilized by researchers for image processing; out of which, NVIDIA Tesla M40 [18,19], NVIDIA GTX 1050 TI [12, 23] NVIDIA GTX 970 [55], and NVIDIA GTX Titan [13, 22, 30, 37, 38, 47] are common.

Tab. 2: Summary of DLNN based skull stripping methods

Author & year	**Methods studied**	**Backbone architecture**	**Hyper-parameters**	**Measures calculated**	**Methods compared**	**Data type**	**System used**
Prasad et al. (2014) [43]	Deformable Model	Intensity Analyses and Motor Control	N/A	Jaccard Index .8478; DSC .9175; Hausdorff 36.35; FPE .0253; FNE .1328	HWA; BET and BSE	T1 from ADNI and manually segmented	N/A
Dai et al. (2015) [8]	Mask R-CNN	CFM based on R-CNN; Spatial Pyarmid Pooling and RPN	N/A	Mean IoU for non-CFM 44 and CFM 50 and 50.9 for design A and B respectively	Inter-dataset comparison	PASCAL Context	Nvidia GTX Titan GPU based on the Caffe Library
Ren et al. (2015) [45]	Region Proposal Networks (RPN)	Faster R-CNN	N/A	Mean Average Precision mAP = .788 and .759 for PASCAL VOC 2007 and 2012 respectively	VGG-16 architecture	T1PASCAL VOC 2007; 2012 and COCO	NVIDIA Tesla K40
Kleesiek et al. (2016) [24]	3d CNN	CNN	N/A	Public Data DSC .9530 and Specificity .9936; Brain Tumor Data DSC .9519 and Specificity .9924	BEaST; Robex; BET; 3dSkullStrip; HWA and BSE	T1; T2; FLAIR IBSR; LPBA; OASIS and Non-enhanced Images	N/A

to be continued in the next page

Tab. 2: Summary of DLNN based skull stripping methods

Author & year	**Methods studied**	**Backbone architecture**	**Hyper-parameters**	**Measures calculated**	**Methods compared**	**Data type**	**System used**
Erden et al. (2017) [14]	3D Fully CNN	3D U-Net and Average Pooling	Epochs = 4, Dropout Rate = 0.5, 0.2 Learning Rate = 0.0004 ”	DSC = .71	Intra Algorithm Comparison	285×T1 BraTS Multimodal Brain Tumor Segmentation	N/A
He et al. (2017) [18]	Mask R-CNN	Faster R-CNN; Feature Pyramid NetworkFPN; ResNet-50-C4; ResNet-101-RPN; ResNeXt-101-FPN	Learning Rate = 0.02; Weight Decay 0.0001; Momentum = 0.9	mAP; Average Time = 195 ms/image	Inter-dataset comparison	COCO	Nvidia Tesla M40 GPU x8
Lin et al. (2017a) [32]	Feature Pyramid Networks	Faster R-CNN; ResNet-50; ResNet-101; VGG-16	Learning Rate = 0.02 for first 60k batches; Learning Rate = 0.002 for next 20k batches; RoIs per image = 2000	AR increased by 8 points; AP on COCO and PASCAL increased by 8 and 3 points respectively; Average Spead 0.165 img/sec for ResNet 50; Average Spead 0.19 img/sec for ResNet 101	G-RMI; AttractioNet; Faster R-CNN; Multipath	COCO-2016; PASCAL	NVIDIA M40 GPU

to be continued in the next page

Tab. 2: Summary of DLNN based skull stripping methods

Author & year	**Methods studied**	**Backbone architecture**	**Hyper-parameters**	**Measures calculated**	**Methods compared**	**Data type**	**System used**
Lin et al. (2017b) [31]	Focal Loss and RetinaNet	FPN; ResNet; ConvNet; ReLU	Learning Rate = 0.01; divided by 10 after 60k and then at 80k iterations; Momentum = 0.9; Weight Decay = 0.0001	AP increased by 3 to 4 points; Average Speed of RetinaNet-ResNet-101-600 is 50 ms/per image faster	Two Staged Methods; Faster R-CNN+++; Faster R-CNN – FPN; Faster R-CNN – GRMI; Faster R-CNN – TDM; One Stage Method; YOLOv2; SSD-513; DSSD-513	COCO	NVIDIA M40 GPU
Mehta et al. (2017) [39]	BrainSegNet	CNN	Epochs = 30; Momentum= 0.75; Learning Rate = 0.05	mDC = 0.844 for IBSR; 0.824 for LONI-LPBA; 0.840 for Hammers67; 0.808 for Hammers83	MALP; Patch based; Classification based	T1 from LONI-LPBA; Hammers-67; Hammers-83; IBSR; MICCAI	NVIDIA Tesla K40
Dey and Hong (2018) [10]	CompNet	CNN	Learning Rate = 0.001; Epochs = 10	DSC = .9827; Sensitivity = .9826; Specificity = .9980	Plain U-Net; Dense U-Net; Probability CompNet; Plain CompNet	T1 from OASIS	N/A

to be continued in the next page

Tab. 2: Summary of DLNN based skull stripping methods

Author & year	**Methods studied**	**Backbone architecture**	**Hyper-parameters**	**Measures calculated**	**Methods compared**	**Data type**	**System used**
Eide (2018) [13]	3D U-Net; DeepMedic; 3D CNN by Kleesiek et al. [24]	ReLU	Learning Rate = 0.0005; iEpochs = 35	DSC; Senstivity; Specificity	Inter-method and Inter-data comparison	T1 from OASIS; LPBA; Data provided by St. Olavs Hospital and LiTS for liver images	Nvidia TitanX
Hu et al. (2018) [19]	Transfer Learning in Mask R-CNN	Faster R-CNN; ResNet-50-FPN; ResNet-101-FPN	Learning Rate = 0.02x 0.1 after 60k and 80k iterations; Weight Decay = 0.0001; Momentum = 0.9	mAP; IoU threshold = 0.5 - 0.95	Inter-dataset comparison	COCO; PASCAL VOC; Visual Genome	Nvidia Tesla M40 GPU x8
Lucena et al. (2018) [36]	Simultaneous Truth and Performance Level Estimation	2D FCN U-Net; Loss Function = Negative of DSC	Learning Rate = 0.00001	DSC = .9579; Sensitivity = .9771; Specificity = .9887; Hausdorff = 1.49; Mean = 0.075	Gold Standard only	LONI-LPBA; CC-359; and OASIS	N/A
Roy and Maji (2018) [49]	Anatomic Region of Surgical Interest ARoSI	Rough Fuzzy Connectedness	N/A	DSC = .9625 and .9553; Senstivity = .9555 and .9520; Specificity = .9965 and .9985 for BrainWeb and NAMIC respectively	AFNI; BSE; BET; ROBEX; BEaST; CNN	T1 from BrainWeb; NAMIC with Normal Control; and NAMIC with Chizophrenic	N/A

to be continued in the next page

Tab. 2: Summary of DLNN based skull stripping methods

Author & year	**Methods studied**	**Backbone architecture**	**Hyper-parameters**	**Measures calculated**	**Methods compared**	**Data type**	**System used**
Roy et al. (2018) [47]	CNN	AlexNet; Gaussian Filter to the binary segmentation	N/A	DSC = .9719; Jaccard Index = .9454; PPV = .9963; Volume Difference = .0477	BEaST; Spectre; Robex; manual method	T1 from MPRAGE	Nvidia Titan X GPU
Selvathi and Vanmathi (2018) [52]	CNN	ReLU; NonLocal Mean for noise reduction	N/A	Senstivity ¿ .87; Specificity ¿ .94; Accuracy ¿ .918	Intra-method and Inter-image comparison	OASIS	N/A
Valvano et al. (2018) [55]	CNN	U-Net – CNN; ReLU – CNN	2x CNN layers of 3x3; 1 stroid in first and 2 stroids in second layer	DSC = 0.965; FNR = 0.2; FPR = 0.8	Robex; FSL; BSE; AFNI; ANTS	T1 from NFBS	Intel Xeon E5-2620 v4 CPU; GPU NVIDIA GTX 970;10 GHz; with 16 cores and 32 threads
Wang et al. (2018) [58]	Non-local Neural Networks	Mask R-CNN; ResNet; Batch-Norm; ImageNet	Learning Rate = 0.01; Momentum = 0.9; Wieght Decay = 0.0001	Average Precision	Inter-dataset comparison	Kinetics dataset; 246k and 20k videos; Charades dataset; 8k and 8k	N/A

to be continued in the next page

Tab. 2: Summary of DLNN based skull stripping methods

Author & year	**Methods studied**	**Backbone architecture**	**Hyper-parameters**	**Measures calculated**	**Methods compared**	**Data type**	**System used**
Yilmaz et al. (2018) [61]	Multi-stable Cellular Neural Network – MCNN	Contrast Enhancement Using Linear Image Combinations Algorithm CEULICA	N/A	Jaccard = 0.838; DSC = 0.899; True Positive = 0.151; True Negative = 0.013	BET; BSE	T1 from BrainWeb; MIDAS-NAMIC; Healthy PeopleHospital	Intel Core TM i7-382060GHZ processor; 16 GB RAM and 64 bit OS
Dai et al. (2019) [7]	Transfer Learning and Multi Output Net MO-Net	3D U-Net; Two staged training	N/A	DSC on MALC = .785; DSC on HAA = 0.843	U-Net FS; U-Net FT; SLAN T8; SLAN T27	UKBB; HAA; MICCAI; MALC	NVIDIA GPU
Dalca et al. (2019) [9]	Atlas Based CNN	CNN	N/A	DSC = .835	Inter-dataset comparison	T1 from OASIS; ABIDE; ADHD; MCIC; PPMI; HABS; and Harvard GSP. PD and manuualy annotated images	NVIDIA Titan Xp GPU.

to be continued in the next page

Tab. 2: Summary of DLNN based skull stripping methods

Author & year	**Methods studied**	**Backbone architecture**	**Hyper-parameters**	**Measures calculated**	**Methods compared**	**Data type**	**System used**
Hwang et al. (2019) [21]	3D U-Net	Extension of 2D U-Net; ReLU; Max Pooling; Batch Normalization	N/A	DSC = .9903; Senstivity = .9853; Specificity = .9953	BSC; ROBEX; Kleesiek's DLNN	T1 from NFBS	2x NVIDIA 1080Ti GPU
Isensee et al. (2019) [22]	HD-BET	U-Net – CNN	N/A	DSC 0.976; Hausdorff Distance 3	Robex; BET; BSE; 3dSkull-Stripping; ANTS	EORTC; LPBA NFBS; Calgary-Compinas	NVIDIA TITAN Xp GPU
Lucena et al. (2019) [37]	Simultaneous Truth and Performance Level Estimation	2D U-Net; reffered as CONSNet	Learning Rate = 0.001; Exponential Decay = 0.995 after each epoch; Fixed Kernel Size = 3 x 3	DSC = .9718; Senstivity = .9891; Specificity = .9946; Hausdorff Distance = 713; Symmetric Surface to Surface Mean Distance SSSMD = .037	ANT's; BEaST; BET; BSE; HWA; MBWSS; OPTIBET; ROBEX; STAPLE-12	T1 from Calgary-Campinas; LPBA; OASIS	NVIDIA had 12 Gbyte; CPU Xeon E3-1220 v3; 4x 10 GHz Intel; GeForce Titan X

to be continued in the next page

Tab. 2: Summary of DLNN based skull stripping methods

Author & year	**Methods studied**	**Backbone architecture**	**Hyper-parameters**	**Measures calculated**	**Methods compared**	**Data type**	**System used**
Wang and Li (2019) [57]	Label Fusion Method	Pixel Grayscale Probability Information; Sparse Representation; Weighted Voting	N/A	DSC; Recall; Precision; Hausdorff Distance HD	FreeSurfer; Lin et al. method [34]; Kushibar et al. method [26]; Xu et al. method [60]; Liu et al. method [35]	T1 from IBSR	PC Intel Core 70 GHz CPU
Dutta et al. (2020) [12]	Time-Distributed U-Net based CNN TD-U Net-CNN	U-Net – CNN; and ReLU	N/A	Model Accuracy = 5 3%	Intra-method	T1w; T2w; FLAIR from MICCAI BraTS; MICCAI BraTS	NVidia GTX 1050Ti GPU; 4 GB of VRAM; 100 GB of DRAM; an Intel i5-8600 K CPU overclocked to 10 GHz
Kim et al. (2020) [23]	Cascade 3D U-Net	U-Net; Active Learning; CNN	Learning Rate = 10 - 5; Weight Decay = 0.0005; Momentum = 0.9; Epochs = 300	Root Mean Squre RMS = 0.86	CNN; Manually Segemeted	MR images of kidney	NVIDIA GTX 1080 TI

to be continued in the next page

Tab. 2: Summary of DLNN based skull stripping methods

Author & year	**Methods studied**	**Backbone architecture**	**Hyper-parameters**	**Measures calculated**	**Methods compared**	**Data type**	**System used**
Kuang et al. (2020) [25]	Skull R-CNN	Faster R-CNN; Skeleton-based region proposal method	RMSprop with = 0.9; Learning Rate = 0.0001	Average Precision AP = 0.62	Faster R-CNN + FPN; Skull R-CNN + FPN	N/A	NVIDIA GTX 1080 TI
Li et al. (2020) [30]	Anatomical Context-Encoding Network – ACENet	2D CNN; U-Net; Spatial Squeeze and Excitations	Weight Decay = 0.0001; Momentum = 0.9; Dropout Rate = 0.1; Epochs = 100	Average Time to Segment 10 seconds; i DSC 0.8	Intra-dataset Comparison	T1 from MALC; ADNI; Mindboggle-101; SchizBull	NVIDIA TITAN XP GPU; 12GB of RAM
Manjón et al. (2020) [38]	Deep Intracranial Cavity Extraction – DeepICE	ReLU	Epochs = 20	DSC = .9889	Non-Local Intracranial Cavity Extraction – NICE; BEaST; VBM8	T1 from OASIS; BSTP	Nvidia TitanX 12 GB RAM

3. Research Gap

In the light of intensive literature review, we have come to the conclusion that the most recent development has been made in the domain of DLNN and the scientific progress has led the experts of digital image processing to successfully experiment with the latest and robust CNN variant named as Mask R-CNN [18] for image segmentation. The comprehensive literature audit did not provide sufficient empirical evidence pertaining to the use of Mask R-CNN for skull stripping. The availability of deep learning weights for hundreds of objects and classes and non-availability of the same for the skull stripping in giant public digital libraries like COCO etc. are also empirical evidences addressing the dearth of research stated above in the realm of image segmentation. The research gap identified and discussed above needs prompt attention of researchers. Therefore, the scientific research study may be carried out to experiment skull stripping using Mask R-CNN along with its underlying structure and auxiliaries to ultimately bridge the existing research gap.

References

[1] J. Bernsen. Dynamic thresholding of gray level images. In *Proc. 8th Int. Conf. on Pattern Recognition ICPR*, page 1251–1255, Paris, France, 27–31 Oct 1986.

[2] A. S. Bhadauria, V. Bhateja, M. Nigam, and A. Arya. Skull stripping of brain MRI using mathematical morphology. In S. Satapathy et al., editors, *Smart Intelligent Computing and Applications, Proc. 3rd Int. Conf. Smart Computing and Informatics SCI 2018-19 (Vol. 1)*, volume 159 of *Smart Innovation, Systems and Technologies*, pages 775–780, Bhubaneswar, India, 21–22 Dec 2018. Springer, Singapore 2020. doi:10.1007/978-981-13-9282-5_75.

[3] P.-F. Chen, R. G. Steen, A. Yezzi, and H. Krim. Brain MRI T1-map and T1-weighted image segmentation in a variational framework. In *Proc. 2009 IEEE Int. Conf. Acoustics, Speech and Signal Processing*, pages 417–420, Taipei, Taiwan, 19-24 Apr 2009. doi:10.1109/ICASSP.2009.4959609.

[4] M. Cheour. Advantages of brain MRI. *RadiologyInfo.org*, 2010.

[5] C. A. Cocosco, V. Kollokian, R. K.-S. Kwan, and A. C. Evans. *BrainWeb*: Simulated Brain Database, 1998. `https://brainweb.bic.mni.mcgill.ca`. [Accessed 10 Oct 2020].

[6] D. L. Collins, A. P. Zijdenbos, V. Kollokian, J. G. Sled, N. J. Kabani, C. J. Holmes, and A. C. Evans. Design and construction of a realistic digital brain phantom. *IEEE Transactions on Medical Imaging*, 17(3):463–468, 1998. doi:10.1109/42.712135.

[7] C. Dai, Y. Mo, E. Angelini, Y. Guo, and W. Bai. Transfer learning from partial annotations for whole brain segmentation. In *Domain Adaptation and Representation Transfer and Medical Image Learning with Less Labels and Imperfect Data. Proc. MICCAI Workshop on Domain Adaptation and Representation Transfer DART 2019*, volume 11795 of *Lecture Notes in Computer Science*, pages 199–206, Shenzen, China, 13 Oct 2019. doi:10.1007/978-3-030-33391-1_23.

[8] J. Dai, K. He, and J. Sun. Convolutional feature masking for joint object and stuff segmentation. In *Proc. 2015 IEEE Conf. Computer Vision and Pattern Recognition CVPR*, pages 3992–4000, Boston, MA, USA, 7-12 Jun 2015. doi:10.1109/CVPR.2015.7299025.

[9] A. V. Dalca, E. Yu, P. Golland, et al. Unsupervised deep learning for bayesian brain MRI segmentation. In D. Shen, T. Liu, T. M. Peters, et al., editors, *Medical Image Computing and Computer Assisted Intervention – MICCAI 2019*, volume 11766 of *Lecture Notes in Computer Science*, pages 356–365, 2019. doi:10.1007/978-3-030-32248-9_40.

[10] R. Dey and Y. Hong. CompNet: Complementary segmentation network for brain MRI extraction. In *Proc. Int. Conf. Medical Image Computing and Computer-Assisted Intervention MICCAI 2018*, volume 11072 of *Lecture Notes in Computer Science*, pages 628–636, Granada, Spain, 16-20 Sep 2018. doi:10.1007/978-3-030-00931-1_72.

[11] J. Doshi, G. Erus, Y. Ou, et al. Multi-atlas skull-stripping. *Academic Radiology*, 20(12):1566–1576, 2013. doi:10.1016/j.acra.2013.09.010.

[12] J. Dutta, D. Chakraborty, and D. Mondal. Multimodal segmentation of brain tumours in volumetric MRI scans of the brain using time-distributed U-Net. In A. K. Das et al., editors, *Proc. Conf. Computational Intelligence in Pattern Recognition CIPR 2019*, volume 999 of *Advances in Intelligent Systems and Computing*, pages 715–725, 2020. doi:10.1007/978-981-13-9042-5_62.

[13] Ø. A. Eide. Skull stripping MRI images of the brain using deep learning. Master's thesis, Norwegian University of Science and Technology, 2018. `https://ntnuopen.ntnu.no/ntnu-xmlui/handle/11250/2566509`.

[14] B. Erden, N. Gamboa, and S. Wood. 3D convolutional neural network for brain tumor segmentation. Technical report, Computer Science, Stanford University, Stanford, USA, 2017. `http://cs231n.stanford.edu/reports/2017/pdfs/526.pdf`.

[15] M. Everingham, L. van Gool, C. Williams, et al. The PASCAL Visual Object Classes homepage, 2012. `http://host.robots.ox.ac.uk/pascal/VOC/`. [Accessed 10 Oct 2020].

[16] A. Fatima, A. R. Shahid, B. Raza, et al. State-of-the-art traditional to the machine-and deep-learning-based skull stripping techniques, models, and algorithms. *Journal of Digital Imaging*, 33(6):1–22, 2020. doi:10.1007/s10278-020-00367-5.

[17] F. J. Galdames, F. Jaillet, and C. A. Perez. An accurate skull stripping method based on simplex meshes and histogram analysis for magnetic resonance images. *Journal of Neuroscience Methods*, 206(2):103–119, 2012. doi:10.1016/j.jneumeth.2012.02.017.

[18] K. He, G. Gkioxari, P. Dollár, and R. Girshick. Mask R-CNN. In *Proc. IEEE Int. Conf. Computer Vision ICCV*, pages 2961–2969, Venice, Italy, 22-29 Oct 2017. doi:10.1109/ICCV.2017.322.

[19] R. Hu, P. Dollár, K. He, et al. Learning to segment every thing. In *Proc. 2018 IEEE Conf. Computer Vision and Pattern Recognition CVPR*, pages 4233–4241, Salt Lake City, USA, 18-23 Jun 2018. doi:10.1109/CVPR.2018.00445.

[20] Y. Huang and L. C. Parra. Fully automated whole-head segmentation with improved smoothness and continuity, with theory reviewed. *PloS ONE*, 10(5):e0125477, 2015. doi:10.1371/journal.pone.0125477.

[21] H. Hwang, H. Z. U. Rehman, and S. Lee. 3D U-Net for skull stripping in brain MRI. *Applied Sciences*, 9(3):569, 2019. doi:10.3390/app9030569.

[22] F. Isensee, M. Schell, I. Pflueger, et al. Automated brain extraction of multisequence MRI using artificial neural networks. *Human Brain Mapping*, 40(17):4952–4964, 2019. doi:10.1002/hbm.24750.

[23] T. Kim, K. Lee, S. Ham, et al. Active learning for accuracy enhancement of semantic segmentation with CNN-corrected label curations: Evaluation on kidney segmentation in abdominal CT. *Scientific Reports*, 10:366, 2020. doi:10.1038/s41598-019-57242-9.

[24] J. Kleesiek, G. Urban, A. Hubert, et al. Deep MRI brain extraction: A 3D convolutional neural network for skull stripping. *NeuroImage*, 129:460–469, 2016. doi:10.1016/j.neuroimage.2016.01.024.

[25] Z. Kuang, X. Deng, L. Yu, et al. Skull R-CNN: A CNN-based network for the skull fracture detection. In T. Arbel et al., editors, *Proc. 3rd Conf. Medical Imaging with Deep Learning MIDL*, volume 121 of *Proceedings of Machine Learning Research*, pages 382–392, Montreal, Canada, 06-08 Jul 2020. `http://proceedings.mlr.press/v121/kuang20a.html`.

[26] K. Kushibar, S. Valverde, S. González-Villà, et al. Automated sub-cortical brain structure segmentation combining spatial and deep convolutional features. *Medical Image Analysis*, 48:177–186, 2018. doi:10.1016/j.media.2018.06.006.

[27] P. J. LaMontagne, T. L. S. Benzinger, J. C. Morris, et al. Oasis-3: Longitudinal neuroimaging, clinical, and cognitive dataset for normal aging and Alzheimer disease. *medRxiv*, 2019. doi:10.1101/2019.12.13.19014902.

[28] P. J. LaMontagne, T. L. S. Benzinger, J. C. Morris, et al. OASIS Open Access Series of Imaging Studies, 2019. `https://www.oasis-brains.org`. [Accessed 10 Oct 2020].

[29] K. Landheer, R. F. Schulte, M. S. Treacy, et al. Theoretical description of modern ^{1}H in Vivo magnetic resonance spectroscopic pulse sequences. *Journal of Magnetic Resonance Imaging*, 51(4):1008–1029, 2020. doi:10.1002/jmri.26846.

[30] Y. Li, H. Li, and Y. Fan. ACEnet: Anatomical context-encoding network for neuroanatomy segmentation. *arXiv*, 2020. arXiv:2002.05773 [eess.IV]. `https://arxiv.org/abs/2002.05773`.

[31] T. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár. Focal loss for dense object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(2):318–327, 2020. doi:10.1109/TPAMI.2018.2858826.

[32] T.-Y. Lin, P. Dollár, R. Girshick, et al. Feature pyramid networks for object detection. In *Proc. 2017 IEEE Conf. Computer Vision and Pattern Recognition CVPR*, pages 936–944, Honolulu, Hawaii, 22-25 Jul 2017. doi:10.1109/CVPR.2017.106.

[33] T.-Y. Lin, G. Patterson, M. R. Ronchi, et al. COCO. Common Objects in Context, 2020. `https://cocodataset.org`. [Accessed 10 Oct 2020].

[34] X.-B. Lin, X.-X. Li, and D.-M. Guo. Registration error and intensity similarity based label fusion for segmentation. *IRBM*, 40(2):78–85, 2019. doi:10.1016/j.irbm.2019.02.001.

[35] Y. Liu, Y. Wei, and C. Wang. Subcortical brain segmentation based on atlas registration and linearized kernel sparse representative classifier. *IEEE Access*, 7:31547–31557, 2019. doi:10.1109/ACCESS.2019.2902463.

[36] O. Lucena, R. Souza, L. Rittner, et al. Silver standard masks for data augmentation applied to deep-learning-based skull-stripping. In *Proc. 2018 IEEE 15th International Symposium on Biomedical Imaging ISBI*, pages 1114–1117, Washington, USA, 4-7 Apr 2018. doi:10.1109/ISBI.2018.8363766.

[37] O. Lucena, R. Souza, L. Rittner, et al. Convolutional neural networks for skull-stripping in brain MR imaging using silver standard masks. *Artificial Intelligence in Medicine*, 98:48–58, 2019. doi:10.1016/j.artmed.2019.06.008.

[38] J. V. Manjón, J. E. Romero, R. Vivo-Hernando, et al. Deep ICE: A deep learning approach for MRI intracranial cavity extraction. *arXiv*, 2020. arXiv:2001.05720 [q-bio.QM]. `https://arxiv.org/abs/2001.05720`.

[39] R. Mehta, A. Majumdar, and J. Sivaswamy. BrainSegNet: a convolutional neural network architecture for automated segmentation of human brain structures. *Journal of Medical Imaging*, 4(2):1–11, 2017. doi:10.1117/1.JMI.4.2.024003.

[40] S. Moldovanu, L. Moraru, and A. Biswas. Robust skull-stripping segmentation based on irrational mask for magnetic resonance brain images. *Journal of Digital Imaging*, 28(6):738–747, 2015. doi:10.1007/s10278-015-9776-6.

[41] W. Niblack. *An Introduction to Digital Image Processing*. Prentice Hall, 1986.

[42] M. Otsu. A threshold selection method from gray-level histograms. *IEEE Trans. Systems, Man and Cybernetics*, 9(1):62–66, 1979. doi:10.1109/TSMC.1979.4310076.

[43] G. Prasad, A. A. Joshi, A. Feng, et al. Skull-stripping with machine learning deformable organisms. *Journal of Neuroscience Methods*, 236:114–124, 2014. doi:10.1016/j.jneumeth.2014.07.023.

[44] H. Z. U. Rehman, H. Hwang, and S. Lee. Conventional and deep learning methods for skull stripping in brain MRI. *Applied Sciences*, 10(5):1773, 2020. doi:10.3390/app10051773.

[45] S. Ren, K. He, R. B. Girshick, and J. Sun. Faster R-CNN: towards real-time object detection with region proposal networks. *arXiv*, 2015. arXiv:1506.01497 [cs.CV]. `http://arxiv.org/abs/1506.01497`.

[46] S. Ren, K. He, R. B. Girshick, and J. Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(6):1137–1149, 2017. doi:10.1109/TPAMI.2016.2577031.

[47] S. Roy, A. Knutsen, A. Korotcov, et al. A deep learning framework for brain extraction in humans and animals with traumatic brain injury. In *Proc. 2018 IEEE 15th International Symposium on Biomedical Imaging ISBI*, pages 687–691, Washington, USA, 4-7 Apr 2018. doi:10.1109/ISBI.2018.8363667.

[48] S. Roy and P. Maji. A simple skull stripping algorithm for brain MRI. In *Proc. 2015 8th Int. Conf. Advances in Pattern Recognition ICAPR*, pages 1–6, Kolkata, India, 4-7 Jan 2015. doi:10.1109/ICAPR.2015.7050671.

[49] S. Roy and P. Maji. An accurate and robust skull stripping method for 3-D magnetic resonance brain images. *Magnetic Resonance Imaging*, 54:46–57, 2018. doi:10.1016/j.mri.2018.07.014.

[50] G. Ruffini, M. D. Fox, O. Ripolles, et al. Optimization of multifocal transcranial current stimulation for weighted cortical pattern targeting from realistic modeling of electric fields. *Neuroimage*, 89:216–225, 2014. doi:10.1016/j.neuroimage.2013.12.002.

[51] J. Sauvola and M. Pietikäinen. Adaptive document image binarization. *Pattern Recognition*, 33(2):225–236, 2000. doi:10.1016/S0031-3203(99)00055-2.

[52] D. Selvathi and T. Vanmathi. Brain region segmentation using convolutional neural network. In *2018 4th Int. Conf. Electrical Energy Systems ICEES*, pages 661–666, Chennai, India, 7-9 Feb 2018. doi:10.1109/ICEES.2018.8442394.

[53] H. Tariq, A. Muqeet, A. Burney, Akhtar H. M., and H. Azam. Otsu's segmentation: Review, visualization and analysis in context of axial brain MR slices. *Journal of Theoretical & Applied Information Technology*, 95(22), 2017. `http://www.jatit.org/volumes/Vol95No22/9Vol95No22.pdf`.

[54] H. Tariq and M. Shahbaz. MAFA: Multispectral adaptive fuzzy algorithm for edge detection on MRI of head scan. *International Journal of Computer Applications*, 182(48):49–54, 2019. doi:10.5120/IJCA2019918737.

[55] G. Valvano, N. Martini, A. Leo, et al. Training of a skull-stripping neural network with efficient data augmentation. *arXiv*, 2018. arXiv:1810.10853 [cs.CV]. `https://arxiv.org/abs/1810.10853`.

[56] A. van der Plas. MRI techniques, 2016. `https://www.startradiology.com/the-basics/mri-technique/`. [Accessed 10 Oct 2020].

[57] M. Wang and P. Li. Label fusion method combining pixel greyscale probability for brain MR segmentation. *Scientific Reports*, 9:17987, 2019. doi:10.1038/s41598-019-54527-x.

[58] X. Wang, R. Girshick, A. Gupta, and K. He. Non-local neural networks. In *Proc. 2018 IEEE Conf. Computer Vision and Pattern Recognition CVPR*, pages 7794–7803, Salt Lake City, USA, 18-23 Jun 2018. doi:10.1109/CVPR.2018.00813.

[59] A. Worth, C. Haselgrove, and D. Kennedy. IBSR. The Internet Brain Segmentation Repository, 2007. `https://www.nitrc.org/projects/ibsr/`. [Accessed 10 Oct 2020].

[60] L. Xu, H. Liu, E. Song, et al. Automatic labeling of MR brain images through extensible learning and atlas forests. *Medical Physics*, 44(12):6329–6340, 2017. doi:10.1002/mp.12591.

[61] B. Yilmaz, A. Durdu, and G. D. Emlik. A new method for skull stripping in brain MRI using multistable cellular neural networks. *Neural Computing and Applications*, 29(8):79–95, 2018. doi:10.1007/s00521-016-2834-2.

[62] J. Zhou, H.-Y. Heo, L. Knutsson, et al. APT-weighted MRI: Techniques, current neuro applications, and challenging issues. *Journal of Magnetic Resonance Imaging*, 50(2):347–364, 2019. doi:10.1002/jmri.26645.

Multi-View Attention-Based Late Fusion (MVALF) CADx System for Breast Cancer Using Deep Learning

Hina Iftikhar[1,2], Ahmad Raza Shahid[1,2], Basit Raza[1,2],
Hasan Nasir Khan[1,2]
[1]*Medical Imaging and Diagnostics Laboratory (MID),*
National Centre of Artificial Intelligence (NCAI), Islamabad, Pakistan
[2]*Department of Computer Science, COMSATS University Islamabad (CUI), Pakistan*
basit.raza@comsats.edu.pk

Abstract. Breast cancer is a leading cause of death among women. Early detection can significantly reduce the mortality rate among women and improve their prognosis. Mammography is the first line procedure for early diagnosis. In the early era, conventional Computer-Aided Diagnosis (CADx) systems for breast lesion diagnosis were based on just single view information. The last decade evidence the use of two views mammogram: Medio-Lateral Oblique (MLO) and Cranio-Caudal (CC) view for the CADx systems. Most recent studies show the effectiveness of four views of mammogram to train CADx system with feature fusion strategy for classification task. In this paper, we proposed an end-to-end Multi-View Attention-based Late Fusion (MVALF) CADx system that fused the obtained predictions of four view models, which is trained for each view separately. These separate models have different predictive ability for each class. The appropriate fusion of multi-view models can achieve better diagnosis performance. So, it is necessary to assign the proper weights to the multi-view classification models. To resolve this issue, attention-based weighting mechanism is adopted to assign the proper weights to trained models for fusion strategy. The proposed methodology is used for the classification of mammogram into normal, mass, calcification, malignant masses and benign masses. The publicly available datasets CBIS-DDSM and mini-MIAS are used for the experimentation. The results show that our proposed system achieved 0.996 AUC for normal vs. abnormal, 0.922 for mass vs. calcification and 0.896 for malignant vs. benign masses. Superior results are seen for the classification of malignant vs benign masses with our proposed approach, which is higher than the results using single view, two views and four views early fusion-based systems. The overall results of each level show the potential of multi-view late fusion with transfer learning in the diagnosis of breast cancer.

Key words: breast cancer, mammogram, four-view mammogram, information fusion, late fusion, transfer learning.

1. Introduction

Breast cancer is one of the most death-causing invasive diseases among women. In 2018, 2.1 million cases of breast cancer were recorded by the World Health Organization (WHO) and 627 000 women died of breast cancer, which is 6.5% of all cancer-related deaths in that year [49]. The death rate has been decreasing since the last few decades. The decrease is due to the advancement in early diagnosis, treatment, and awareness about the symptoms [37]. However, in the past years women death rate was still high due to the diagnosis is frequently still too late. Early diagnosis prevents the patient

from invasive tumor and it also increases the survival rate by five to ten years. Mammography is a reliable and initial diagnostic method for early diagnosis of breast cancer. Mammograms are low energy X-rays of the breast and radiologist use it to identify the abnormalities in the breast. Breast screening has been performed on two views: Cranio-Caudal (CC) and Medio-lateral Oblique (MLO) of the left and right breast. CC view is top-down screening and MLO view is taken under 45 degrees [19, 46].

Breast cancer includes calcifications and masses. Calcifications are the deposits of calcium in woman's breast and can be shown clearly as white dots in the screening process. There are further two types of calcification: macrocalcifications and microcalcifications [29]. Macrocalcifications are large white spots that are considered as the non-cancerous and are dispersed randomly in the breast. Microcalcifications are the small white deposits of calcium and are mostly considered as non-cancerous. Although, if these deposits are clustered together then this may be alarming as early breast cancer [47]. Masses are the lesions in woman's breast that can be cancerous or non-cancerous. The benign masses, that is, the non-cancerous ones are smooth or oval in shape with circumscribed boundary. The masses that are known as cancerous, that is, malignant, spread into their neighborhood by forming spicules. Diagnosis of masses is a challenging task due to the variations in their shape, appearance and size [29]. However, manual detection of the symptoms of cancer using mammograms is susceptible to human errors and laborious due to variability. In the current technical era, Computer-Aided Diagnosis (CADx) systems are used for reliable and fast diagnosis of disease. CADx systems have potential to reduce the heavy workload of the radiologist. These systems served as a second reader to improve the accuracy of the final decision.

In the last few years, deep learning has become one of the most successful methods in computer vision tasks [25]. Especially, Convolution Neural Networks (CNNs) have been proved as the reason for the boom of deep learning. Deep learning-based CADx systems [11, 13, 36] have attained the level appropriate for producing more realistic solutions in tumor diagnosis. The four major steps are involved in CNN-based CADx systems to assist the radiologist in making the final decision [50]. Firstly, the preprocessing step is performed to remove the noise from images. In the second step the region of the tumor is segmented out from the image. The feature extraction task is carried out for the region of the tumor in the third step. In the last step, the tumor classification task is performed. Traditional CADx systems were based on manual handcrafted features, which have shown the limited accuracy for complex problems. Several studies have been performed to build a CADx system for breast lesion classification and detection. In 2013, Kozegar et al. [27] used the traditional feature selection and machine learning techniques for iterative breast segmentation. Their proposed system had the ability to classify the segmented region of the lesion. Other results and the literature on the segmentation-based mammography analysis systems can be found for example in [7].

A number of recent studies have been published on fully automated CNN-based

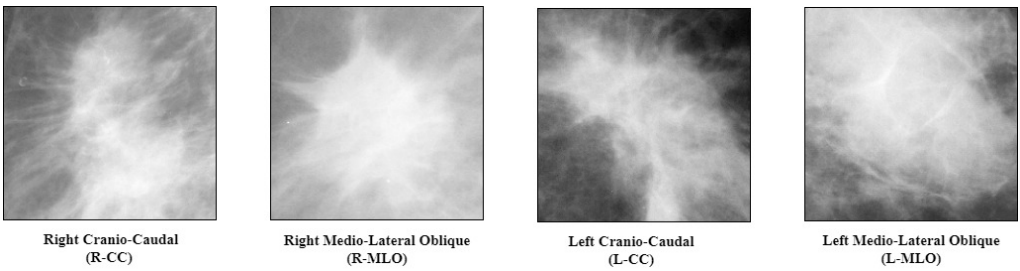


Fig. 1. Examples of ROIs of four mammographic views in the CBIS-DDSM dataset.

CADx system for tumor detection and classification tasks [8, 11, 13, 22]. The deep learning-based CADx systems have been introduced for different medical domains, for example brain tumor detection, lung disease diagnosis, lymph node, breast cancer diagnosis, and many others. We mainly focused on breast lesion classification [4, 5, 8, 10, 18, 32, 33, 34]. CNN is an end-to-end supervised learning process without any descriptor on the whole raw image. CNN learns the discriminant features automatically and its most surprising characteristic is that it achieves good generalization for vision tasks with the 2D input images [29].

Deep CNNs are more complex architectures than CNNs and require a large amount of data to train a model. Due to high computation complexity, training the model on a small amount of data leads to overfit. To overcome this problem, the transfer learning is used. Transfer learning is a technique of transferring the knowledge from one domain to another domain. In medical imaging, where small datasets are available, transferring of knowledge from another domain has been very effective. The knowledge transfer consists in using a network which is pre-trained on images coming from some domain. There are two modes of transferring the knowledge: first, transferring the knowledge from the medical domain, and second, transferring the knowledge from some other domain, for example, the domain of natural images. The current evidences show the high performance of using pre-trained models to achieve better accuracy [12, 22, 29, 36]. In recent years, the authors achieved reasonable accuracies for breast cancer detection and classification task using the transfer learning techniques [2, 12, 29].

Information extracted from multi-view images is more significant for decision making than that extracted from a single view. Multi-view mammograms are used by the radiologist to make a final decision. We will overcome the problem of not gaining profit from the multi-view nature of mammograms in CC and MLO views. In the previous studies, most of the research has been based on single-view images in the development of a CADx system. Breast screening provides the four views: Right MLO (R-MLO), Left MLO (L-MLO), Right CC (R-CC) and Left CC (L-CC) of mammograms as shown in Fig.1. Radiologists always start from the CC view, and when they find any abnormalities

in this view they check the information from all views for making a final decision. Most of the studies focused on the CADx systems based on just two views (CC and MLO mammograms) [8,9]. Recent studies focused on the four-views information-based CADx systems which achieved the best accuracy for breast lesion classification. Multi-view information fusion mainly focuses on the analysis of mammograms using CC and MLO views of the left and right breast. Information fusion is based on two strategies: early fusion and late fusion. Early fusion is used to fuse the extracted features of different models and late fusion is based on combining the results of classification of the multiple models. The results of two-views CC and MLO models are fused to classify the breast lesion into malignant and benign [17] and produce significant results in terms of accuracy for the classification task. In the recent study, Khan et al. proposed a Multi-View Feature Fusion (MVFF) based CADx system that includes three stages [26].

Nevertheless, the multi-view information fusion has gained more success in recent years in context of breast cancer. According to the previous studies on the mammographic views, the breast screening is performed on bilateral view, CC and MLO, of right and left breast. Bassett et al. [6] believed that the CC view, with particular emphasis on the medial view imaging, conveys the most significant information. The CC view is the medial view in screening and has a great aspect of deep tissues to be visualized. Normally, these deep tissues in medial aspect of breast are not possible to capture in the MLO lateral view [6, 19, 45]. However, both projections are complementary to capture the most accurate information. In current era, one of the key challenges is to overcome the high False Positive Rate (FPR) that existed in the previous CADx systems. The four-view fusion systems reduce the high FPR [24]. Wei et al., in 2011, proposed a computer-aided detection system of four view information fusion for mass detection [48]. In comparison with single-view their system performed better in terms of accuracy and FPR. In 2015, Yanfeng Li et al. [30] proposed a bilateral image analysis scheme for mass detection to reduce the FPR. The results show the significance of proposed system in which the approach of bilateral analysis for mass detection reduce the FPR. Among the methods of breast mass detection [30, 31, 48] few of the research works on the multi-view information fusion for classification task [41, 51] use the multi-agent and feature fusion approach, respectively. The results show that the decision fusion mechanism reduces the problem for the classification task. Since the many masses are difficult to identify in one view and give more information in the other view, the late fusion approach reduces the FPR [51]. The four-view information fusion-based CADx systems can be considered as the simulation of radiologist's interpretation and are able to serve as a second reader.

The main focus of this research is to utilize the effectiveness of attention-based weighted late fusion in CADx systems to reduce the false positive rate for mammogram classification. In the late fusion, separate deep CNN models are trained for each view, i.e., L-CC, R-CC, L-MLO, and R-MLO of mammograms. The pre-trained CNN

architectures are used to fine-tune on mammograms to classify the breast lesions. The obtained results of trained models are fused to achieve the best performance in terms of classification of breast masses. The proposed Multi-View Attention-based Late Fusion (MVALF) model outperforms the multi-view model and provides the state-of-the-art technique for mass classification tasks. Our proposed system is evaluated on benchmark dataset CBIS-DDSM (references will be given in Subsection 3.1). The main contributions of this research are as follows.

- A novel attention-based weighting algorithm is proposed to increase the effectiveness of our multi-view late fusion-based CADx system. Each model has its own predictive ability, therefore assigning the equal weights to all the models is not a good approach. In this regard, attention-based weighting algorithm assigns the higher weights to those models which have higher sensitivity.
- A Multi-View Attention-based Late Fusion (MVALF) system is proposed for the diagnosis of breast cancer. The main contribution of this work is to efficiently take the advantage of the four mammographic views of each patient because conventionally developed CADx systems have used two views information and ignored the importance of late fusion of separately trained multi-view models. The proposed MVALF approach yields good performance measures and shows the effectiveness of late fusion for four-view models to reduce the false positive rate.
- The end-to-end system is proposed, which is not limited to just classify the mammogram into cancerous or non-cancerous. The proposed MVALF-based CADx has the ability to classify the mammogram at different levels. The first level is about the classification into normal and abnormal. At the second level, the mammograms are classified on the basis of their abnormality. Finally, at the last level the mammograms are classified according to their level of pathology.

This paper proceeds as follows: Section 2 presents the literature review, Section 3 describes the methodology, Section 4 gives the details of experimentations and the results are discussed in it, and finally Section 5 concludes the paper.

2. Literature review

Many studies have been published on CNN-based CADx systems for breast cancer classification. Chakraborty et al. [10] proposed a novel method that was used to detect non-palpable breast cancer. The automatic diagnosis is difficult due to variability in size, irregularities in shape and occlusions in breast tissue. The proposed method classifies the masses along with characterized oriented tissue and multi-resolution features using Gray-Level Co-Occurrence Matrix (GLCM) and Angle Co-Occurrence Matrix (ACM). Recently Ribli et al. used fast Region-based CNN (R-CNN) for mass detection and classification into malignant and benign [34]. They achieved state-of-the-art performance on the INBreast dataset and their system reached high sensitivity with few false negatives,

and with AUC of 0.85. Al-masni et al. [4] in 2018 proposed a YOLO-based CADx system for breast cancer detection. Their CADx system detects the location and diagnoses the masses and classifies them into benign and malignant class using CNN. The last fully connected layer of architecture is trained on ROI-based mammograms. In 2017, Lotter et al. [32] proposed a methodology for breast cancer mass detection and segmentation. The author proposed a patch-based CNN classifier for lesion classification and achieved 0.92 AUC. In another study, Akselrod-Ballin et al. [3] used fast R-CNN to detect the breast abnormalities on the INBreast dataset and achieved TPR 0.93 and FPI 0.56 for mass mammograms.

Chougrad et al. [12] explored the importance of a pre-trained model and determined the best strategy to train CNNs architectures. They focused on the use of the pre-trained model for classification of breast lesions. The pre-trained models VGG16, ResNet50 and InceptionV3were used instead of random initialization. The proposed full framework for breast cancer screening achieved AUC of 0.9 for masses classification into benign and malignant. Recently, in 2019 Hua Li et al. [29] proposed an improved DenseNet for mammogram classification into benign and malignant class based on a deep learning pre-trained model. The proposed model, DenseNet II, performs the classification task accurately and effectively. AlexNet, VGGNet, GoogleNet, DenseNet and the proposed DenseNet II were trained on processed data. The authors claimed that the system was robust and good at generalization. In the same year, Agarwal et al. [2] proposed a patch-based CNN for automated mass detection. The transfer learning models (ResNet50, VGG16, Inception) were used to train on the CBIS-DDSM dataset and the evaluation revealed that InceptionV3 performed the best on automatic mass detection. The evaluation results demonstrated that patch-based transfer learning CNNs performed substantially well for mass detection on CBIS-DDSM.

While the previous networks were trained on a single view and two views of mammograms, recent years witnessed great advancement in multi-view information-based CADx systems and information fusion of different models attained the state-of-the-art performance [1]. Carneiro et al. proposed a multi-view based CADx system for breast cancer risk prediction using two views of mammograms [8, 9]. Tan et al. proposed a four-view based feature fusion model for near term breast cancer risk prediction [43]. Jiao et al. [23] created and trained a CNN-based CADx system by combining the results of two classifiers and classified the mass mammograms into malignant and benign. They concluded that the results obtained from multi-view model fusion achieved higher classification performance than that using a single view. A similar work has been proposed in 2019, Khan et al. used the early fusion strategy to diagnose the tumor in breast. They utilized the extracted mammographic information of four views. The system had the capability to classify the tumor into malignant and benign. They achieved the classification accuracy of 77% and AUC of 0.84 [26].

In our work, we focus on the attention-based weighted late fusion technique by utilizing the four views of mammogram.

3. Materials and methods

In this section, we first describe the publicly available datasets, data pre-processing, data augmentation, CNN architectures used for our proposed system, evaluation metrics for testing the performance of CADx system, and the overall methodology with attention-based weighting algorithm.

3.1. Dataset

In this study, the dataset that we used to perform the experiments on our proposed MVALF based CADx system were CBIS-DDSM and mini-MIAS. DDSM [20,21] was the first version of CBIS-DDSM. It contains the digital images of mammographic screening of 2 620 patients. It contains the verified pathology information (benign and malignant) of each case. The four view information for each case is available with MLO and CC views of the left and right breasts. CBIS-DDSM [39, 44] is a subset of images selected from the original dataset and curated by expert radiologists [15,28]. It has been used for the training and also for performance evaluation of the proposed MVALF system. The images are compressed and converted into DICOM format. The Mammographic Image Analysis Society (MIAS) is another curated digital mammographic dataset of breast lesions [40] with images of resolution 1024×1024 pixels. The analysis is performed on extracted ROI images of 224×224 pixels of mini-MIAS [14] for normal class. Table 1 shows the detailed description of the train and test split of mammographic dataset using four views.

3.2. Data pre-processing

In order to enhance the performance of the CADx system, we need to perform some mandatory task to make the data clarity better for training a model. We used the ROI-based mammograms from the publicly available dataset. We also performed image

Tab. 1. Dataset description of mammograms in CBIS-DDSM and mini-MIAS.

Abnormality Type	Training	Testing	Total
Normal	3008	512	3520
Abnormal	2864	12	3376
Calcification	1546	256	1802
Mass	1318	256	1574

pre-processing such as contrast and brightness enhancement, resizing and image normalization on the selected datasets. The pre-processing helps to achieve better classification accuracy.

3.3. Data augmentation

Deep learning models perform better when we have a large amount of data. The data in medical imaging domain are very limited in size. The scarcity of the dataset in training the deep learning models leads them to overfit. Data enhancement or data augmentation is an approach to help increase dataset size. It also leads to better robustness and helps to prevent overfitting when training is done on a smaller dataset. We performed data enhancement on our dataset to improve the performance of the system. The images were augmented by rotating by a 0-45 degree angle, the shearing in the range of 0.2, zooming in the range of 0.2, horizontal shifting in the range of 0.2 of the image width, and vertical shifting in the range of 0.2 of the image height. The horizontal flip and vertical flip were performed, and to fill newly created pixels the fill mode strategy was applied. The augmented images were different from each other and there was no exact copy of any of the original images.

3.4. CNN architectures

CNNs are trained on images to recognize the visual pattern with minimal preprocessing. We analyzed the well-known transfer learning models on ImageNet (natural images) [16] along with fine-tuned layers on mammograms. The ImageNet is a dataset containing millions of natural images. ImageNet Large Scale Visual Recognition Challenge (ILSVRC) is a competition for classification and object detection held every year [1, 16]. We have evaluated the performance in the classification of mammograms of the three well known CNN architectures that have been the winners of ILSVRC.

3.4.1. VGGNet

Simonyan et al. in Visual Geometry Group (VGG) from University of Oxford proposed VGGNet [38]. It was much deeper than the previous networks. They used the filter size of 3×3 instead of 5×5, 7×7 or 11×11, as in AlexNet [35]. The network was runner-up of ILSVRC 2015 challenge for image classification with top five error rate of 7.3% and it also performed best in the image localization task. There are many versions of VGGNet; however, VGG16 and VGG19 are the most popular. VGG19 performed better than VGG16 although it is computationally more expensive.

3.4.2. InceptionV3

GoogLeNet was the winner of ILSVRC in 2014 for image classification with top five error rate of 6.7%. Szegedy et al. [42] from Google designed a much deeper network with

22 layers. A novel element known as the inception module was introduced to reduce the computational complexity of the network. In this network the number of parameters was reduced from 60 million (AlexNet) to 4 million.

3.4.3. ResNet50

Residual block network won the ILSVRC 2015 with 3.6% error rate [35]. It is a much deeper network than others with 152 layers. It consists of a residual block where each block contains two 3×3 convolution layers. Skip connections are used in ResNet to remove the vanishing gradient problem [25]. ResNet50 achieved good performance in all tasks such as localization, classification and object detection in ILSVRC.

3.5. Performance Evaluation

The CADx system is evaluated for the correct classification of mammograms. The model is evaluated using sensitivity, specificity, and accuracy as the measures of classification quality. Sensitivity is the True Positive Rate (TPR) and specificity is the True Negative Rate (TNR). Accuracy is measured by the performance of the model in terms of general correctness. We also evaluated the model using the ROC curve and the Area Under the ROC Curve (AUC). ROC curve is a 2-axis presentation with sensitivity on the y-axis and False Positive Rate (FPR) on the x-axis that is calculated as $1 -$ specificity. In the following Equations (1) to (3), sensitivity, specificity and accuracy are calculated in terms of the numbers of True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN) classifications.

$$\text{Sensitivity} = \text{TPR} = \frac{\text{TP}}{\text{TP} + \text{FN}} \,, \tag{1}$$

$$\text{Specificity} = \text{TNR} = \frac{\text{TN}}{\text{TN} + \text{FP}} \,, \tag{2}$$

$$\text{Accuracy} = \text{ACC} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FN} + \text{FP} + \text{TN}} \,. \tag{3}$$

3.6. Proposed four-view model fusion

A fully automated deep CNN-based framework is proposed for mammogram classification using Regions of Interest (ROI's) as input images. Firstly, the dataset is divided into four views: L-CC, R-CC, L-MLO, and R-MLO. Afterwards, the four models are trained on each view separately for all patients. The obtained results from four models of all views are combined to generate the final prediction for mass classification. The prediction fusion of multiple models is known as late fusion [24].

We applied the late fusion strategy on the trained model of each view to generate the final decision. The radiologists also examined the mammograms in the same manner to

make the final decision about the abnormalities. We performed many experiments with variations in hyperparameters. The experiments were made with four view-based CADx systems with various pre-trained models along with the fine-tuning strategy. Fig. 2 shows the proposed MVALF based CADx system for the breast mammogram classification.

3.6.1. Network training

The first stage of the proposed system is related to the model training. At this stage, we fine-tuned the deep CNN models for each view, i.e. L-CC, R-CC, L-MLO, and R-MLO, separately. The best fitted fine-tuned layers have been selected after performing various experiments using different numbers of freezing layers. We also performed experiments for two-view and multi-view cases using pre-trained models. Finally, we concluded from the results that the pre-trained models performed better on multi-view information while the number of datasets was limited. It can be observed that the transferring of knowledge from one domain to another domain helps to achieve better accuracy.

3.6.2. Multiview late fusion strategy

The last level of our system represents the fusion of four view results, which were obtained from the model training phase of each view separately. In breast cancer the screening mammograms are taken from two angles: MLO and CC of left and right breasts. The radiologist makes a final decision after viewing the information from four views. Our proposed CADx system is capable of classifying the mammograms using the four views. Afterwards, the results of all models are fused using the attention-based weighted late fusion strategy and the final decision of the diagnostic task is achieved. The details of the personalized weighting algorithm to prioritize the models are discussed in the next paragraphs.

Attention-based weighting algorithm After training the M models (where $M = 4$) on the four views of mammogram, they have the ability to classify the unseen data into the respective binary classes. Their output is fused to make the final decision. Rather than considering the information of all views equally, the Attention based Weighting Algorithm (AWA) has been adopted. It calculates the weights of predictive score for each view of the models based on their sensitivity to increase the TPR and decrease the FPR.

Let $\text{model}_1, \text{model}_2, \text{model}_3, \ldots, \text{model}_M$ be the M models and $R_1, R_2, R_3, \ldots, R_n$ be the classification results, each of the specific model. Suppose that C is the number of classes of the given dataset labelled as $\text{class}_1, \text{class}_2, \text{class}_3, \ldots, \text{class}_C$. The matrix $W = (w_m)$, $1 \leq m \leq M$ is the weight matrix of M models. The testing image is classified by assigning the label of the model according to the highest score.

In our proposed framework, W is calculated based on TPR. According to the previous studies on the mammogram views, the breast screening is performed on bilateral view

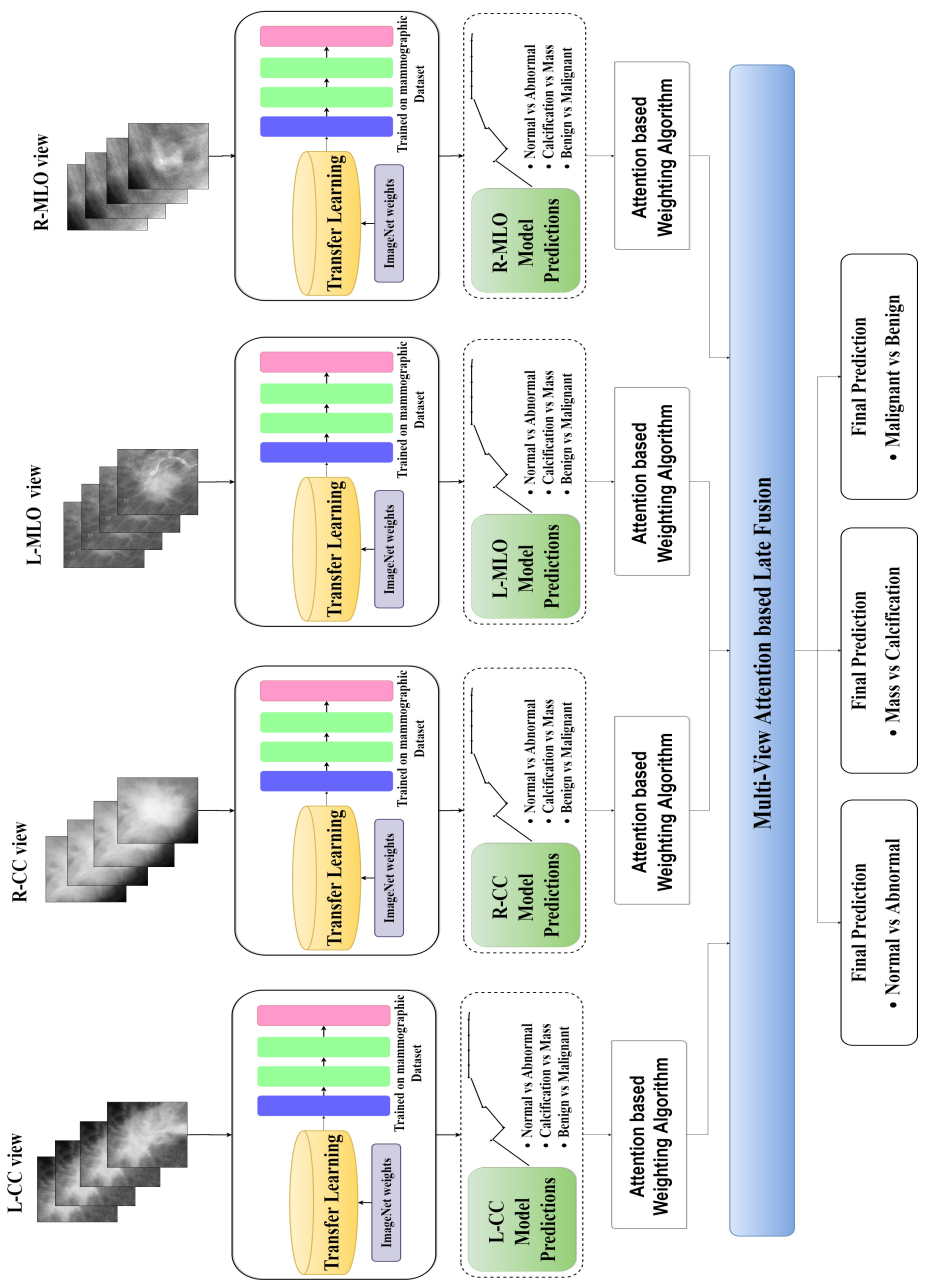


Fig. 2. Proposed Multi-View Attention based Late Fusion (MVALF) CADx system using pre-trained model.

CC and MLO of right and left breast. However, both projections are important to capture the more accurate information. The highest weight is assigned to the view with the highest sensitivity. In our case, the total number of classes is $C = 2$ and the number of models is $M = 4$. The pseudo code for our AWA is presented in Algorithm 3.1.

Algorithm 3.1 Attention-based Weighting Algorithm

for $\text{model}_m \leftarrow \text{model}_1$ **to** model_M **do**
  $\text{TruePos}(m) \leftarrow \langle\text{number of true positive instances in}\rangle(\text{model}_m)$
  $\text{FalseNeg}(m) \leftarrow \langle\text{number of false negative instances in}\rangle(\text{model}_m)$
  $\text{SensitivityM}(m) \leftarrow \text{TruePos}(m)/(\text{TruePos}(m) + \text{FalseNeg}(m))$
  $W(m)_{\text{sen}} \leftarrow \text{SensitivityM}(m)$
end for

4. Results and discussion

In this study, we used the attention-based late fusion strategy and evaluated the different CNN architectures for the classification of mammograms into three levels: mammogram classification, abnormality classification, and pathology classification. Furthermore, we performed experiments on a single view, two views and four views with the early fusion strategy for the comparative study with our proposed CADx system.

4.1. Experimental setup

In the experimental environment, the input size of the ROI image was 224×224. The ROI-based images were pre-processed before training on the CNN architectures. We used the stochastic gradient descent optimization algorithm with 0.0001 learning rate with a momentum of 0.9. The categorical-cross entropy was used as the loss function and the batch size was set between 20 to 50 for training. The dataset had a split of 0.2 for the validation set to evaluate the performance of the correct classification of mammograms. We used the experimental setup for training our models with the specification of NVIDIA Tesla P100, 16 gigabytes of memory, CUDA 10.1 version, Keras 2.2.5 version with TensorFlow 1.15.0 at the backend. The stopping criteria for training the model was set to 200 epochs with the patience level of 15.

4.2. Transfer learning and fine tuning

The transfer learning technique is used in our proposed methodology with fine-tuning strategy. The state-of-the-art pre-trained models (i.e. VGGNet, GoogleNet, ResNet) were trained on the public dataset of ImageNet that contains the natural images of 1000 classes. We removed the last fully connected classification layer of the pre-trained

Tab. 2. The total number of parameters that need to be trained on mammograms using CNN models.

CNN Models	Total Layers	Freezing Layer	Trainable Parameters	Batch Size
VGG19	22	14	14 158 848	50
InceptionV3	311	170	16 338 816	50
ResNet50	175	100	19 452 928	50

models and added two fully connected layers. The first layer has 300 connections and the second layer is used for final classification with two neurons. The approach of freezing layers in the pre-trained model reduces the number of trainable parameters. This helps overcome the problem of computational complexity in deep CNN models. The last, fully connected layers that are fine-tuned on mammograms surpass the overfitting which occurs due to random initialization in deep CNN networks.

The Table 2 shows the total number of layers, freezing layers of pre-trained models, total number of trainable parameters and batch size which was used in our experiments.

4.3. Monitoring the performance of our model

The basic structure of our proposed model is shown in Fig. 2. Our proposed MVALF based CADx system classifies the mammograms at three levels. The first level presents the classification of normal and abnormal mammograms. The second level describes the classification of abnormality into calcification and mass classes. The last level is about the classification of pathology into malignant and benign classes.

4.3.1. Classification into Normal and Abnormal

In the first level, classification of *Normal* and *Abnormal* classes is performed using the proposed MVALF based CADx system. The MVALF based CADx system outperformed the single view, two views and four views-based early fusion. Table 3 shows the performance of the proposed model. The model achieved a good balance between TPR and FPR. The use of transfer learning improves the performance of the proposed system. The four-view models use the weighted information fusion strategy on the basis of TPR, that helps to achieve the AUC of 0.996 shown in Fig. 3. Our proposed MVALF performed better on all the pre-trained models. InceptionV3 and ResNet50 performs slightly better with respect to VGG19. The achievements of the proposed model in comparison to previous studies are shown in Table 6. The proposed MVALF based CADx system performs 7% better than multi-view, two-view and single-view feature fusion.

4.3.2. Classification into Mass and Calcification

Secondly, experiments were performed to classify the abnormality into *Calcifications* and *Masses*. The experimental results in Table 4 show the preformance of the proposed

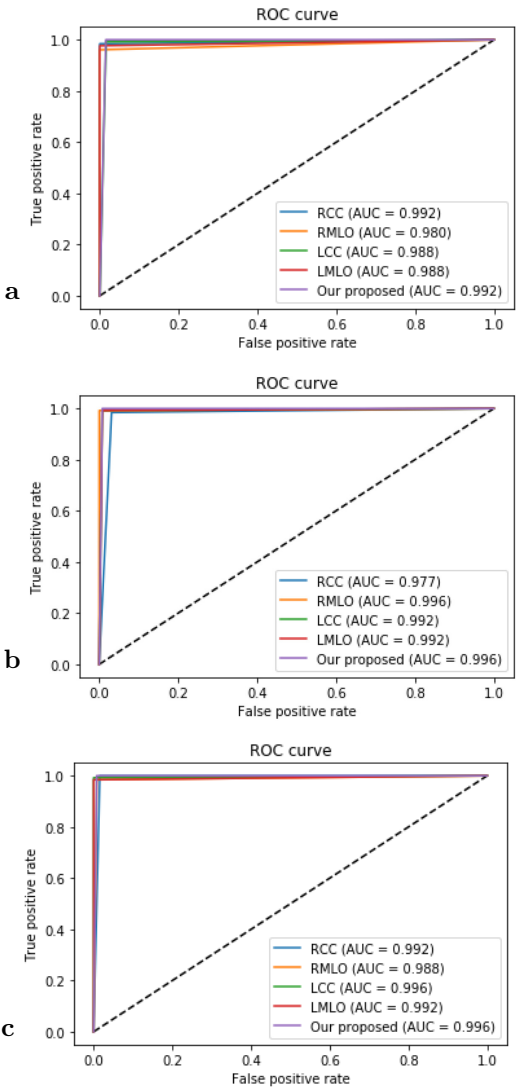


Fig. 3. ROC plotting for *Normal* and *Abnormal* classification. The testing performance of (**a**) VGG19, (**b**) InceptionV3, and (**c**) ResNet50 is presented, using the proposed MVALF-based CADx system.

Tab. 3. Performance measures of proposed MVALF for the classification of *Normal* vs. *Abnormal* mammograms.

Models	Views	Training Accuracy	Testing Accuracy	Sensitivity	Specificity	AUC
VGG19	R-CC	96.5%±0.88%	99.22%±0.68%	98.46%	100%	0.992
	L-CC	99.33%±0.57%	98.83%±0.22%	99.21%	98.45%	0.988
	L-MLO	99.00%±0.98%	98.83%±0.90%	97.71%	100%	0.989
	R-MLO	99.54%±0.22%	98.05%±0.90%	96.24%	100.00%	0.981
	Proposed Multiview (Late Fusion)	–	99.22%±0.78%	100%	98.44%	0.992
InceptionV3	R-CC	98.01%±1.2%	97.66%±1.50%	98.41%	96.92%	0.977
	L-CC	99.26%±0.53%	99.22%±0.71%	99.22%	99.22%	0.992
	L-MLO	99.93%±0.07%	99.22%±0.41%	99.22%	99.22%	0.992
	R-MLO	99.99%±0.10%	99.61%±0.59%	99.22%	99.00%	0.996
	Proposed Multiview (Late Fusion)	–	99.61%±0.29%	100%	99.22%	0.996
ResNet50	R-CC	98.45%±1.50%	99.22%±0.11%	100%	98.46%	0.992
	L-CC	97.44%±2.10%	99.61%±0.30%	99.22%	100%	0.996
	L-MLO	99.56%±0.15%	99.22%±0.13%	98.46%	100%	0.992
	R-MLO	98.28%±1.17%	98.83%±1.23%	98.45%	99.21%	0.988
	Proposed Multiview (Late Fusion)	–	**99.61%±100%**	**100%**	**99.22%**	**0.996**

MVALF-based CADx system. The late fusion of four-view models with their attentional mechanism VGG19 performs better with our proposed late fusion strategy in terms of AUC. However, the MVALF model achieved higher specificity with InceptionV3 in contrast with low sensitivity as compared to VGG19. The main reason behind the best performance of VGG19 for abnormality classification is the good quality of models for each view, i.e. R-CC, L-CC, L-MLO and R-MLO. The weights are assigned on the basis of sensitivity, as each separate model in VGG19 has high sensitivity, so that the model with higher weights improves the overall performance of the system. The model achieves the AUC of 0.922, testing accuracy of 92.12%, sensitivity of 93.55%, and specificity of 90.91%. Fig. 4 shows the ROC curvec of VGG19, InceptionV3 and ResNet50, and as it is clearly shown in the figure, this ensemble of the weighted information of all the views leads to achieving good performance in terms of AUC. The comparison study of the proposed model and the previous approach is shown in Table 6. This study shows the clear difference between the impact of different transfer learning models. The depth of each model has a different impact on the results of the classification task. The VGG19 with very few trainable parameters has achieved good accuracy and AUC for the abnormality classification.

4.3.3. Classification into Malignant and Benign

We performed different experiments for the two-class classification into *Benign* masses and *Malignant* masses. Table 5 shows the different experimental results of each view separately and for our proposed MVALF-based CADx system. The proposed system performed best for the classification task and achieved AUC of 0.896, testing accuracy of

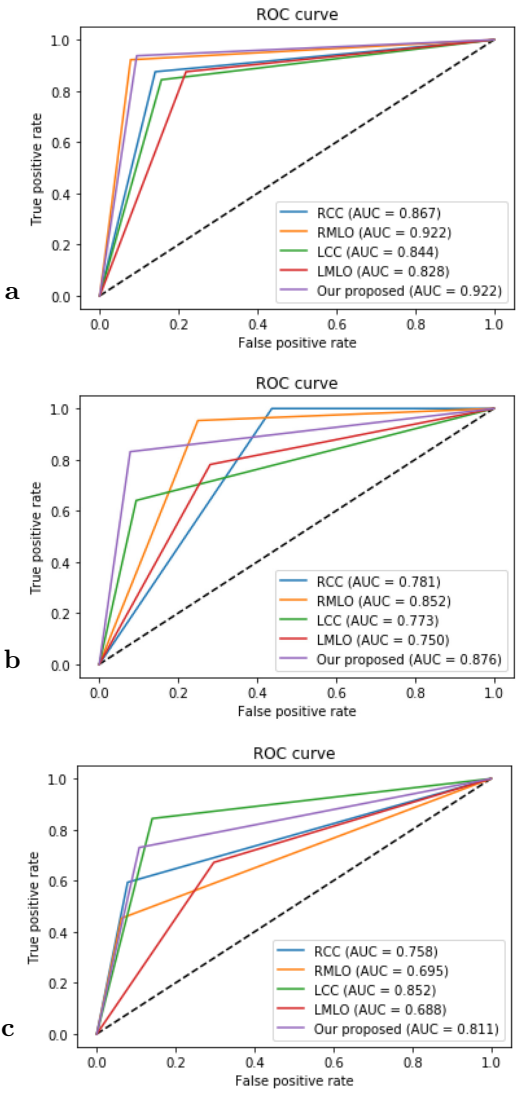


Fig. 4. ROC plotting for *Calcification* and *Mass* classification. The testing performance of (**a**) VGG19, (**b**) InceptionV3, and (**c**) ResNet50 is presented, using the proposed MVALF based CADx system.

Tab. 4. Performance measures of proposed MVALF for the classification of *Mass* vs. *Calcification* mammograms.

Models	Views	Training Accuracy	Testing Accuracy	Sensitivity	Specificity	AUC
VGG19	R-CC	95.09%±1.53%	86.72%±.23%	87.30%	86.15%	0.867
	L-CC	87.79%±1.98%	84.38%±1.57%	84.38%	84.38%	0.844
	L-MLO	88.67%±1.53%	82.81%±1.98%	83.21%	80.00%	0.828
	R-MLO	97.89%±0.98%	92.19%±0.54%	82.19%	82.19%	0.922
	Proposed Multiview (Late Fusion)	–	92.19%±1.56%	93.55%	90.91%	0.922
InceptionV3	R-CC	89.26%±1.98%	78.13%±2.14%	100%	69.57%	0.781
	L-CC	79.10%±2.34%	77.34%±2.19%	79.60%	87.23%	0.773
	L-MLO	79.93%±2.19%	75.00%±2.78%	76.67%	73.53%	0.750
	R-MLO	89.02%±1.78%	85.61%±1.78%	84.12%	79.22%	0.852
	Proposed Multiview (Late Fusion)	–	86.72%±1.57%	78.33%	94.12%	0.876
ResNet50	R-CC	86.45%±0.98%	75.78%±0.98%	69.41%	88.37%	0.758
	L-CC	87.44%±0.97%	85.16%±1.65%	84.62%	85.71%	0.852
	L-MLO	77.09%±2.19%	68.75%±2.45%	68.18%	69.35%	0.688
	R-MLO	78.21%±1.57%	69.53%±2.98%	63.16%	87.88%	0.695
	Proposed Multiview (Late Fusion)	–	**81.25%±1.54%**	**77.33%**	**86.79%**	**0.811**

89.91%, the sensitivity of 86.71%, and the specificity of 94.39%. The performance of our system in term of the ROC curve is shown in Fig. 5. Furthermore, for the comparative study we also performed experiments with single view, two views and four views feature fusion for the mass classification. The results presented in the Table 6 show that our proposed MVALF-based system outperformed and was able to surpass the state-of-art multi-view models.

The comparison between three different state-of-the-art pre-trained models are shown in Fig. 5. The pre-trained model VGG19 outperforms InceptionV3 and ResNet50 for the mass classification in MVALF system. However, our proposed system achieved best results with AUC of 0.896 in contrast with single view, two views and four view early fusion based system which have obtained AUC of 0.737, 0.842 and 0.769, respectively. The proposed MVALF model performs 5% better than the multi-view feature fusion model, 5–10% better than the single and two-views models. The MVALF based CADx system provides a benchmark approach of information fusion for classification tasks into the medical field, especially for breast cancer where four-view information of the patient is available. Table 5 depicts the performance measures of our proposed classifier into *Benign* and *Malignant* cases.

4.3.4. Comparison summary of our work with others

The comparison study was performed to evaluate the performance of our proposed MVALF-based CADx system in comparison to previous studies that use the deep CNN models for mammogram classification tasks. For instance, we compared between single view and two views. Furthermore, we compared our proposed system with the recent

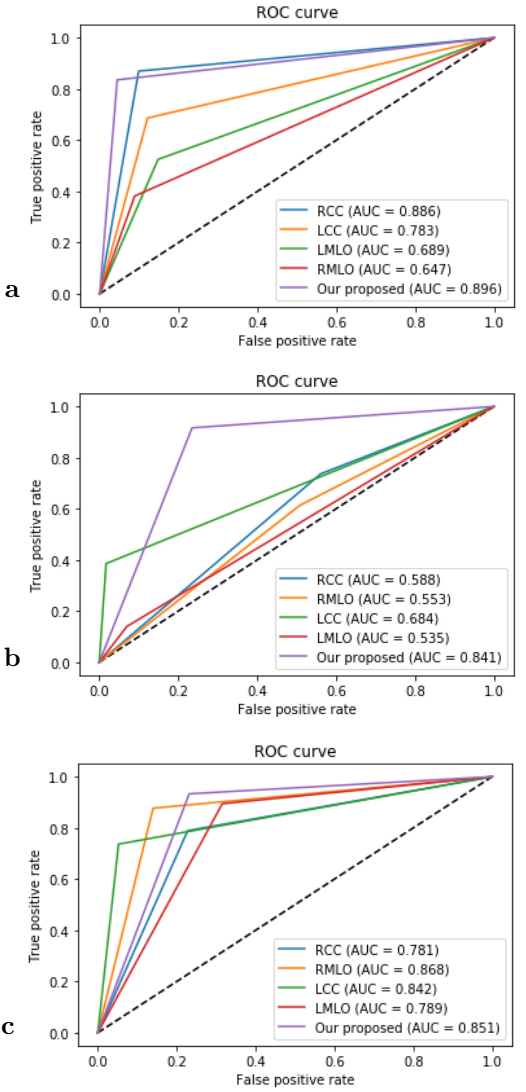


Fig. 5. ROC plotting for *Benign* and *Malignant* classification. The testing performance of (**a**) VGG19, (**b**) InceptionV3, and (**c**) ResNet50 is presented, using the proposed MVALF based CADx system.

Tab. 5. Performance measures of proposed MVALF for the classification of *Malignant mass* vs *Benign mass* mammograms

Models	Views	Training Accuracy	Testing Accuracy	Sensitivity	Specificity	AUC
VGG19	R-CC	96.57%±1.67%	88.64%±1.57%	88.59%	88.71%	0.886
	L-CC	81.33%±0.98%	78.82%±2.18%	75.90%	83.45%	0.783
	L-MLO	88.67%±0.45%	69.81%±2.14%	66.83%	75.89%	0.689
	R-MLO	75.54%±1.56%	66.20%±2.91%	62.35%	79.35%	0.647
	Proposed Multiview (Late Fusion)	–	89.91%±1.57%	86.71%	94.39%	0.896
InceptionV3	R-CC	89.26%±1.98%	77.77%±2.13%	70.50%	79.76%	0.811
	L-CC	73.10%±2.41%	67.40%±1.54%	75.44%	89.00%	0.851
	L-MLO	79.93%±1.58%	70.51%±3.20%	72.96%	77.67%	0.791
	R-MLO	84.02%±1.11%	75.26%±2.91%	73.00%	70.69%	0.785
	Proposed Multiview (Late Fusion)	–	80.07%±2.01%	98.73%	78.43%	0.860
ResNet50	R-CC	86.45%±1.45%	78.07%±2.10%	78.57%	77.59%	0.781
	L-CC	87.44%±1.98%	84.21%±1.98%	78.26%	93.33%	0.868
	L-MLO	77.09%±2.19%	78.95%±2.78%	76.67%	73.91%	0.842
	R-MLO	68.21%±2.98%	76.84%±1.98%	77.50%	86.21%	0.789
	Proposed Multiview (Late Fusion)	–	**83.33%±1.57%**	**94.64%**	**72.41%**	**0.851**

Tab. 6. Comparison with different mammography classification techniques using state-of-the-art pre-trained models on the CBIS-DDSM dataset.

Views	Models	Normal or Abnormal	Mass or Calcification	Malignant or Benign
Single View	VGG19	0.940	0.877	0.737
	InceptionV3	0.907	0.875	0.692
	ResNet50	0.914	0.862	0.644
Two View	VGG19	0.998	0.844	0.843
	InceptionV3	0.938	0.842	0.821
	ResNet50	0.971	0.883	0.811
Four Views (Early Fusion)	Small VGGNet [26]	0.934	0.923	0.769
Proposed Multiview (Late Fusion)	VGG19	0.992	0.922	0.896
	InceptionV3	0.996	0.876	0.860
	ResNet50	0.996	0.811	0.851

study performed on the four-view analysis using feature fusion strategy. Khan et al. in 2019 proposed a small VGGNet with the feature fusion strategy [26]. The system had the capability to classify the breast tumor using mammograms with four views. The results in Table 6 reveal that our proposed MVALF-based CADx system outperform the previous studies. We achieved the AUC of 0.996 for normal and abnormal mammogram classification, AUC of 0.922 for abnormality classification, and AUC of 0.896 for pathology classification.

5. Conclusion

In this work, we proposed a novel multi-view attention-based late fusion CADx system for mammogram classification using the transfer learning approach. We performed experiments using four views information and the results provide the evidence of achieving

the best testing accuracy rate due to late information fusion. We observed that in the late fusion technique for mammogram classification, the overfitting problem occurs due to the unbalance and the limited size of the dataset. According to our assessment, data enhancement plays an important role in reducing the over-fitting problem. Furthermore, the comparison study shows that the proposed model achieves good classification performance and also reduces the computational complexity of the system with the help of the pre-trained model. We conclude that VGGNet pre-trained on ImageNet models with fine-tuning performs the best among all the pre-trained models for our proposed attention-based weighted late fusion approach. Table 6 demonstrates the comparative overview of the previous studies with the proposed MVALF-based CADx system. The results clearly show the effectiveness of the proposed technique. Our system provides a baseline for the new approach to attention-based weighted late fusion using the CBIS-DDSM for abnormality and pathology classification.

In the future work, we will experiment to analyze the impact of different sources for the improvement of the proposed CADx system.

Acknowledgement

This work has been supported by Higher Education Commission under Grant # 2 (1064), and is carried out at the Medical Imaging and Diagnostics (MID) Lab at COMSATS University Islamabad, under the umbrella of the National Center of Artificial Intelligence (NCAI), Pakistan.

References

[1] Large Scale Visual Recognition Challenge 2015 (ILSVRC2015), 2015. `http://www.image-net.org/challenges/LSVRC/2015/results`. [Aceessed Jun 2020].

[2] R. Agarwal, O. Diaz, X. Lladó, et al. Automatic mass detection in mammograms using deep convolutional neural networks. *Journal of Medical Imaging*, 6(3):31409, 2019. doi:10.1117/1.JMI.6.3.031409.

[3] A. Akselrod-Ballin, L. Karlinsky, A. Hazan, et al. Deep learning for automatic detection of abnormal findings in breast mammography. In M. J. Cardoso, T. Arbel, G. Carneiro, et al., editors, *Proc. Int. Workshops on Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support DLMIA, ML-CDS, in conjunction with MICCAI 2017*, volume 10553 of *Lecture Notes in Computer Science*, pages 321–329. Springer, Québec City, QC, Canada, 14 Sep 2017. doi:10.1007/978-3-319-67558-9_37.

[4] M. A. Al-masni, M. A. Al-antari, J.-M. Park, et al. Simultaneous detection and classification of breast masses in digital mammograms via a deep learning YOLO-based CAD system. *Computer methods programs in biomedicine*, 157:85–94, 2018. doi:10.1016/j.cmpb.2018.01.017.

[5] J. Arevalo, F. A. González, Ramos-Pollán, et al. Representation learning for mammography mass lesion classification with convolutional neural networks. *Computer methods programs in biomedicine*, 127:248–257, 2016. doi:10.1016/j.cmpb.2015.12.014.

[6] L. W. Bassett, I. A. Hirbawi, N. DeBruhl, and M. K. Hayes. Mammographic positioning: evaluation from the view box. *Radiology*, 188(3):803–806, 1993. doi:10.1148/radiology.188.3.8351351.

[7] M. Bator and M. Nieniewski. Detection of cancerous masses in mammograms by template matching: Optimization of template brightness distribution by means of evolutionary algorithm. *Journal of Digital Imaging*, 25(1):162–172, 2012. doi:10.1007/s10278-011-9402-1.

[8] G. Carneiro, J. Nascimento, and A. P. Bradley. Unregistered multiview mammogram analysis with pre-trained deep learning models. In N. Navab, J. Hornegger, W. M. Wells, and A. F. Frangi, editors, *Proc. Int. Conf. Medical Image Computing and Computer-Assisted Intervention MICCAI 2015*, volume 9351 of *Lecture Notes in Computer Science*, pages 652–660, Munich, Germany, 5-9 Oct 2015. Springer. doi:10.1007/978-3-319-24574-4_78.

[9] G. Carneiro, J. Nascimento, and A. P. Bradley. Automated analysis of unregistered multi-view mammograms with deep learning. *IEEE Transactions on Medical Imaging*, 36(11):2355–2365, 2017. doi:10.1109/TMI.2017.2751523.

[10] J. Chakraborty, A. Midya, and R. Rabidas. Computer-aided detection and diagnosis of mammographic masses using multi-resolution analysis of oriented tissue patterns. *Expert Systems with Applications*, 99:168–179, 2018. doi:10.1016/j.eswa.2018.01.010.

[11] H. Chen, D. Ni, J. Qin, et al. Standard plane localization in fetal ultrasound via domain transferred deep neural networks. *IEEE Journal of Biomedical Health Informatics*, 19(5):1627–1636, 2015. doi:10.1109/JBHI.2015.2425041.

[12] H. Chougrad, H. Zouaki, and O. Alheyane. Deep convolutional neural networks for breast cancer screening. *Computer Methods Programs in Biomedicine*, 157:19–30, 2018. doi:10.1016/j.cmpb.2018.01.011.

[13] F. Ciompi, B. de Hoop, S. J. van Riel, et al. Automatic classification of pulmonary peri-fissural nodules in computed tomography using an ensemble of 2D views and a convolutional neural network out-of-the-box. *Medical Image Analysis*, 26(1):195–202, 2015. doi:10.1016/j.media.2015.08.001.

[14] A. F. Clark. The mini-MIAS database of mammograms, 2012. `http://peipa.essex.ac.uk/info/mias.html` [Accessed Jun 2020].

[15] K. Clark, B. Vendt, K. Smith, et al. The cancer imaging archive (TCIA): Maintaining and operating a public information repository. *Journal of Digital Imaging*, 26(6):1045–1057, 2013. doi:10.1007/s10278-013-9622-7.

[16] J. Deng, W. Dong, R. Socher, et al. ImageNet: A large-scale hierarchical image database. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition CVPR 2009*, pages 248–255, Miami, FL, USA, 20-25 Jun 2009. IEEE. doi:10.1109/CVPR.2009.5206848.

[17] S. Dhahbi, W. Barhoumi, and E. Zagrouba. Multi-view score fusion for content-based mammogram retrieval. In A. Verikas, P. Radeva, and D. Nikolaev, editors, *Proc. 8th Int. Conf. Machine Vision ICMV 2015*, volume 9875 of *Proc. SPIE*, page 987515, Barcelona, Spain, 8 Dec 2015. doi:10.1117/12.2228614.

[18] N. Dhungel, G. Carneiro, and A. P. Bradley. A deep learning approach for the analysis of masses in mammograms with minimal user intervention. *Medical Image Analysis*, 37:114–128, 2017. doi:10.1016/j.media.2017.01.009.

[19] G. W. Eklund. The art of mammographic positioning. In M. Friedrich and E. A. Sickles, editors, *Radiological Diagnosis of Breast Diseases*, pages 75–88. Springer, 2000. doi:10.1007/978-3-642-60919-0_6.

[20] M. Heath, D. Bowyer, R. Kopans, et al. The digital data base for screening Mammography. In M. J. Yaffe, editor, *Proc. 5th Int. Workshop on Digital Mammography*, pages 212–218, Toronto, Canada, 11-14 Jun 2000. Medical Physics Publishing, Madison, WI, USA. `http://www.eng.usf.edu/cvprg/Mammography/Database.html`.

[21] M. Heath, K. Bowyer, D. Kopans, et al. Current status of the digital database for screening mammography. In N. Karssemeijer, M. Thijssen, J. Hendriks, and L. van Erning, editors, *Digital Mammography*, pages 457–460. Springer Netherlands, Dordrecht, 1998. doi:10.1007/978-94-011-5318-8_75.

[22] B. Q. Huynh, H. Li, and M. L. Giger. Digital mammographic tumor classification using transfer learning from deep convolutional neural networks. *Journal of Medical Imaging*, 3(3):34501, 2016. doi:10.1117/1.JMI.3.3.034501.

[23] Z. Jiao, X. Gao, Y. Wang, and J. Li. A deep feature based framework for breast masses classification. *Neurocomputing*, 197:221–231, 2016. doi:10.1016/j.neucom.2016.02.060.

[24] A. Jouirou, A. Baâzaoui, and W. Barhoumi. Multi-view information fusion in mammograms: A comprehensive overview. *Information Fusion*, 52:308–321, 2019. doi:10.1016/j.inffus.2019.05.001.

[25] A. Khan, A. Sohail, U. Zahoora, and A. S. Qureshi. A survey of the recent architectures of deep convolutional neural networks. *Artificial Intelligence Review*, 53(8):5455–5516, 2020. doi:10.1007/s10462-020-09825-6.

[26] H. N. Khan, A. R. Shahid, B. Raza, A. H. Dar, and H. Alquhayz. Multi-view feature fusion based four views model for mammogram classification using convolutional neural network. *IEEE Access*, 7:165724–165733, 2019. doi:10.1109/ACCESS.2019.2953318.

[27] E. Kozegar, M. Soryani, B. Minaei, and I. Domingues. Assessment of a novel mass detection algorithm in mammograms. *Journal of Cancer Research and Therapeutics*, 9(4):592, 2013. doi:10.4103/0973-1482.126453.

[28] R. S. Lee, F. Gimenez, A. Hoogi, et al. A curated mammography data set for use in computer-aided detection and diagnosis research. *Scientific Data*, 4:170177, 2017. doi:10.1038/sdata.2017.177.

[29] H. Li, S. Zhuang, D.-A. Li, J. Zhao, and Y. Ma. Benign and malignant classification of mammogram images based on deep learning. *Biomedical Signal Processing Control*, 51:347–354, 2019. doi:10.1016/j.bspc.2019.02.017.

[30] Y. Li, H. Chen, Y. Yang, et al. A bilateral analysis scheme for false positive reduction in mammogram mass detection. *Computers in Biology and Medicine*, 57:84–95, 2015. doi:10.1016/j.compbiomed.2014.12.007.

[31] X. Liu, T. Zhu, L. Zhai, and J. Liu. Improvement of mass detection in mammogram using multi-view information. In C. M. Falco and X. Jiang, editors, *Proc. 8th Int. Conf. Digital Image Processing ICDIP 2016*, volume 10033 of *Proc. SPIE*, page 100334M, Chengdu, China, 29 Aug 2016. doi:10.1117/12.2244627.

[32] W. Lotter, G. Sorensen, and D. Cox. A multi-scale CNN and curriculum learning strategy for mammogram classification. In M. J. Cardoso, T. Arbel, G. Carneiro, et al., editors, *Proc. Int. Workshops on Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support DLMIA, ML-CDS, in conjunction with MICCAI 2017*, volume 10553 of *Lecture Notes in Computer Science*, pages 169–177. Springer, Québec City, QC, Canada, 14 Sep 2017. doi:10.1007/978-3-319-67558-9_20.

[33] W. Peng, R. V. Mayorga, and E. M. A. Hussein. An automated confirmatory system for analysis of mammograms. *Computer Methods Programs in Biomedicine*, 125:134–144, 2016. doi:10.1016/j.cmpb.2015.09.019.

[34] D. Ribli, A. Horváth, Z. Unger, et al. Detecting and classifying lesions in mammograms with deep learning. *Scientific Reports*, 8(1):4165, 2018. doi:10.1038/s41598-018-22437-z.

[35] O. Russakovsky, J. Deng, H. Su, et al. ImageNet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211–252, 2015. doi:10.1007/s11263-015-0816-y.

[36] H.-C. Shin, H. R. Roth, M. Gao, et al. Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning. *IEEE Transactions on Medical Imaging*, 35(5):1285–1298, 2016. doi:10.1109/TMI.2016.2528162.

[37] R. L. Siegel, K. D. Miller, and A. Jemal. Cancer statistics, 2019. *CA: A Cancer Journal for Clinicians*, 69(1):7–34, 2019. doi:10.3322/caac.21551.

[38] K. Simonyan and A Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint*, 2014. arXiv:1409.1556v6.

[39] K. Smith, J. Kirby, D. Runbin, et al. CBIS-DDSM – Curated Breast Imaging Subset of DDSM. In TCIA Team [44]. [Accessed Jun 2020]. `https://wiki.cancerimagingarchive.net/display/Public/CBIS-DDSM`.

[40] J. Suckling, J. Parker, D. Dance, et al. The Mammographic Images Analysis Society digital mammogram database. In A. G. Gale, S. M. Astley, D. R. Dance, and A. Y. Cairns, editors, *Digital Mammography*, volume 1069 of *Exerpta Medica International Congress Series*, pages 375–378. Elsevier, 1994. `http://www.wiau.man.ac.uk/services/MIAS/` [Inoperative].

[41] L. Sun, L. Li, W. Xu, et al. A novel classification scheme for breast masses based on multi-view information fusion. In *Proc. 4th Int. Conf. Bioinformatics and Biomedical Engineering iCBBE 2010*, pages 1–4, Chengdu, China, 18-20 Jun 2010. IEEE. doi:10.1109/iCBBE.2010.5517742.

[42] C. Szegedy, W. Liu, Y. Jia, et al. Going deeper with convolutions. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition CVPR 2015*, pages 1–9, Boston, MA, USA, 7-12 Jun 2015. doi:10.1109/CVPR.2015.7298594. `https://www.cv-foundation.org/openaccess/content_cvpr_2015/html/Szegedy_Going_Deeper_With_2015_CVPR_paper.html`.

[43] M. Tan, J. Pu, S. Cheng, et al. Assessment of a four-view mammographic image feature based fusion model to predict near-term breast cancer risk. *Annals of Biomedical Engineering*, 43(10):2416–2428, 2015. doi:10.1007/s10439-015-1316-5.

[44] TCIA Team, editors. *The Cancer Imaging Archive*, 2021. [Accessed Jun 2020]. `https://www.cancerimagingarchive.net/`.

[45] P. D. Trieu, P. C. Brennan, W. Lee, E. Ryan, et al. The value of the craniocaudal mammographic view in breast cancer detection: a preliminary study. In C. K. Abbey and C. R. Mello-Thoms, editors, *Proc. Conf. SPIE Medical Imaging 2013: Image Perception, Observer Performance, and Technology Assessment*, volume 8673 of *Proc. SPIE*, page 86731J, Lake Buena Vista, FL, United States, 28 Mar 2013. doi:10.1117/12.2006821.

[46] C. J. Vyborny and R. A. Schmidt. Mammography as a radiographic examination: an overview. *RadioGraphics*, 9(4):723–764, 1989. doi:10.1148/radiographics.9.4.2667052.

[47] WebMD. Breast Calcifications, 2020. `https://www.webmd.com/women/guide/breast-calcification-symptoms-causes-treatments`. [Accessed Jun 2020].

[48] J. Wei, H.-P. Chan, C. Zhou, et al. Computer-aided detection of breast masses: Four-view strategy for screening mammography. *Medical Physics*, 38(4):1867–1876, 2011. doi:10.1118/1.3560462.

[49] World Health Organization. Cancer. Facts sheet, 2018. `https://www.who.int/news-room/fact-sheets/detail/cancer`. [Aceessed Jun 2020].

[50] N. I. R. Yassin, S. Omran, E. M. F. El Houby, and H. Allam. Machine learning techniques for breast cancer computer aided diagnosis using different image modalities: A systematic review. *Computer Methods and Programs in Biomedicine*, 156:25–45, 2018. doi:10.1016/j.cmpb.2017.12.012.

[51] H. Zhao, W. Xu, L. Li, and J. Zhang. Classification of breast masses based on multi-view information fusion using multi-agent method. In *Proc. 5th Int. Conf. Bioinformatics and Biomedical Engineering iCBBE 2011*, pages 1–4, Wuhan, China, 10-12 May 2011. IEEE. doi:10.1109/iCBBE.2011.5780304.

Hina Iftikhar completed the M.Sc. Computer Science degree from COMSATS University Islamabad (CUI), Islamabad Pakistan in the area of Medical Image Analysis. Her research interest includes the development of CADx system for early detection of breast cancer and computer vision task using machine learning and deep learning approaches. She has a number of conference papers in international conferences.

Ahmad Raza Shahid is currently working as Assistant Professor at COMSATS University Islamabad (CUI), Islamabad, Pakistan. He did his Ph.D. in Computer Science in York, UK in 2012. During his PhD he worked on automatically building a WordNet for four languages, namely, English, German, French and Greek. After his Ph.D. he has been working in the areas of Computer Vision and Pattern Recognition, Machine Learning, and Natural Language Processing. A few of the problems that he has worked on include cancer detection, pedestrian detection, driver fatigue detection, and data mining.

Basit Raza received his master's degree in computer science from the University of Central Punjab, Lahore, Pakistan. He received his Ph.D. in computer science from International Islamic University Islamabad and University Technology Malaysia in 2014. Currently, he is an Assistant Professor in the Department of Computer Science, COMSATS University Islamabad (CUI), Islamabad, Pakistan. He is member of Medical Imaging and Diagnostics Lab, National Center of Artificial Intelligence (NCAI) since 2018. His research interests are data science, medical imaging, database management systems, data mining, data warehousing, machine learning, deep learning and artificial intelligence. Dr. Raza has authored several papers in refereed journals and has been serving as a reviewer for prestigious journals, such as Applied Soft Computing, Swarm and Evolutionary Computation, Swarm Intelligence, Applied Intelligence, IEEE Access and Future Generation Computer Systems.

Hasan Nasir Khan received the B.Sc. degree in computer science from COMSATS University Islamabad, Sahiwal, Pakistan, in 2016. He pursued the M.Sc. degree in computer science at COMSATS University Islamabad, Islamabad, Pakistan in 2019. He is working as a Research Assistant with the Medical Imaging and Diagnostics Lab at COMSATS University Islamabad, under the umbrella of National Center of Artificial Intelligence, Pakistan. His research interest includes the development of computer-aided diagnosis systems for early diagnosis of breast cancer using artificial intelligence and computer vision techniques. Hasan Nasir Khan was a recipient of the Prime Minister of Pakistan's National ICT Scholarship Award in 2012. He has published 5 international conference proceedings and a journal paper.

Normal Patch Retinex Robust Alghoritm for White Balancing in Digital Microscopy

Radosław Roszczyk[1], Artur Krupa[2], Izabella Antoniuk[2]
[1]*Faculty of Electrical Engineering*
Warsaw University of Technology, Warsaw, Poland
radoslaw.roszczyk@pw.edu.pl
[2]*Institute of Information Technology*
Warsaw University of Life Sciences – SGGW, Warsaw, Poland
artur_krupa@sggw.edu.pl

Abstract. The acquisition of accurately coloured, balanced images in an optical microscope can be a challenge even for experienced microscope operators. This article presents an entirely automatic mechanism for balancing the white level that allows the correction of the microscopic colour images adequately. The results of the algorithm have been confirmed experimentally on a set of two hundred microscopic images. The images contained scans of three microscopic specimens commonly used in pathomorphology. Also, the results achieved were compared with other commonly used white balance algorithms in digital photography. The algorithm applied in this work is more effective than the classical algorithms used in colour photography for microscopic images stained with hematoxylin-phloxine-saffron and for immunohistochemical staining images.

Key words: auto white balance algorithm, microscope image processing, staining of microscopic slides, digital microscopy.

1. Introduction

The consistency of visible colours is one of the fascinating possibilities that the human eye provides. A person can look at an object from any angle, but regardless of the varying lighting conditions, the colour of the observed element will not change significantly. This effect is achievable due to the highly complex structure of the eye. In the case of humans, colour perception is compensated by the adaptive capacity of tissues and the capabilities of the human brain. However, computer algorithms cannot deal with this type of problem. At the same time, colour stability is the essential element in case of image processing and analysis as well as recognition of objects placed in processed scenes. Hence, colour stability plays a significant role, especially in the algorithms of automatic image segmentation [24] or feature extraction [25] in microscopic medical images.

The use of electronic image capture technology in medicine is based on the solution used in the past in conventional photography, i.e. the photosensitive film. The main difference, however, is the colour representation used in traditional and in microscopic photography.

In the case of digital photography, the image is created as a result of interaction of the incident light reflected from the surface of the photographed object with the light sensor. Recording of a microscopic image (referred to as a slide), for which the light source is placed centrally below the object on the laboratory glass, is realised by recording the flux passing through the object and falling on the optical system. As a result of this treatment, the image is more natural than in the case with the incident light and has a white background. In this case, the light penetrates through the microscopic specimen.

The ability to adjust the white balance to the colour space is an important feature in the case of images obtained by optical sensors in modern electron microscopes. Most of the images are represented in the popular RGB space, so the resulting images are also stored in this space.

The eye and the physiological mechanisms which perform the image processing in the human vision system are not fully explored, but the eyes' ability to capture objects in the vicinity of the light beam which is reflected from them, is known. In such a case the brain adjusts the input light spectrum so that the colour perception is consistent with the colour values of the observed object far from the beam. This, in turn, means that despite the different lighting parameters, the objects illuminated in this way remain perceptually uniform.

The research issues discussed are related primarily to microscopic images which are used by a very wide scientific community. Medical images, chemical reagents – the use of microscopy allows researchers to see what cannot be seen with the human eye. The medical images constitute a basis in the research methodology, and have a direct impact on the therapy applied and its effectiveness.

Medical images require proper preparation both before their acquisition and after saving the digital form of the slide taken from the sample. A comparison of two images made under the same laboratory conditions with the same staining may still show some colour discrepancy, as described in [17]. One of the parameters which have the most important influence on the quality of the image is the white balance. An incorrectly performed optimization of the white balance of an image can affect the possibility of further processing of the used material. To our best knowledge, at present in the domain of medical imaging there is a deficit of tools for white balance adjustment.

Unambiguity in medicine is critical, but it often happens that images are created by different medical groups, using completely different devices. Each team has a different approach to the calibration process, which is a condition of operational reliability. Just as the measurement of alcohol in the exhaled air should be performed with a certified and calibrated device, the materials used in the tests should be prepared in an appropriate manner. Despite the existence of various recommendations, the measurement conditions or rules describing the required sequence of actions, the images often remain without suitable colour preprocessing.

In the case of microscopic medical images, they are usually analysed by using feature

Tab. 1. Description of data sets containing test images.

Stain	Quantity	Series	Lens	Magnification
HPS	100	HPS	Hitachi HV F22CL	20x
IHC	50	CK34	CIS VCC-FC60FR19CL	40x
	50	KI68	CIS VCC-FC60FR19CL	20x

extraction for objects present in such images. Inadequately prepared images can hinder the implementation of further operations. In recent years, many methods have been defined to solve this type of problem. In [25] the extraction of features by the use of basic morphological segmentation and the description of the observed objects are shown, defining their geometric parameters along with their variability. However, the method reacts differently to the images illuminated with different intensity. Each time it is necessary to select the appropriate parameters of operation.

The colour information contained in the slides is very useful in the assessment of similarity of regions of the images. The availability of information in three colour channels instead of one grey level makes it possible to apply more advanced methods of analysis. Among such methods we can distinguish the L*a*b* segmentation method [1] or the automatic white balance methods [6]. We can also distinguish normalization methods, such as histogram extension, colour transfer method, or spectral methods [10]. Mainly the latter is often used in microbiology and pathomorphology. It is based on estimating the tinting spectrum by adjusting the proportions of tinting to the intensity range for each pixel, even in the case of significant differences in shades.

2. Images

The experimental part described in the paper was based on the use of microscopic medical images. The collection consisted of 200 images coming from microscopic scans of actual tissues. All the images are 1500×1500 pixels size. The images used were made using two types of staining which enhance important features necessary for medical analysis: hematoxylin-phloxine-saffron staining (HPS) and immunohistochemistry staining (IHC). Staining with IHC was performed with the application of two different biological markers: CK34 and KI67. Sets of images stained with IHC were collected using slides obtained from the Archives of the Military Medical University. The set stained with HPS was prepared from the OpenSlide public microscope slide collection (from [7], part of [9], described in [8]). Images were acquired using a variety of devices and in different optical and colour settings. Table 1 contains information about the number of samples and the file settings used during recording.

The images were divided into four series of 25 images for each type of staining. Each series was taken from different microscope slides, and the images in the series were

selected randomly. This ensured that all the cases considered were independent and that the methods used were not closely correlated with the properties of the chosen data subset.

3. Existing methods

Most of the widely available algorithms designed to ensure colour stability have been successfully implemented in colour photography [13]. However, such solutions have not yet been used for medical applications, and in particular not for the preparation of materials taken from microscopic sources.

The concept of white balance in digital technology is related to certain limitations of optical sensors performing the acquisition operations for the projection of the light beam reflected from the object onto the matrix area. White balance consists in adjusting the colour depending on the ambient light and the light falling onto the optics. Incorrect selection of light balance causes that the object correctly seen by the human eye will be, for example, too much inclined in the direction of *yellow* (giving the impression of warm) or too much biased versus to *blue* (giving the impression of cold).

In its simplest form, this phenomenon is represented by the colour temperature scale (Fig.1). Leaving aside the physical issues of the nature of light, we can see that the determination of the temperature makes it possible to select appropriate parameters of the image colour transformation using this scale. It should be therefore determined whether a given light is *cold* or *warm*. The higher the temperature on the scale, the colder the light, and warmer in the opposite direction. The aforementioned terms of *warm* and *cold* are theoretical concepts that characterise the generally accepted perception of colour by humans.

A typical home light source has a temperature of around 3000 K. Daylight (solar) also called *white* light has the temperature of around 5400 K during the day and 6700 K on a cloudy day. At night, it is almost completely blue, and the temperature ranges from 8000 to 10000 K.

Fig. 1. Illustration of the light colour temperature scale indicating the relationship between the temperature (referenced to an ideal black-body radiator) and the colour of the light source perceived in a given range – warm for lower temperatures and cold for higher temperatures (saturation is amplified to make hues visible). Source: [11], used in [26], among others. See also [27].

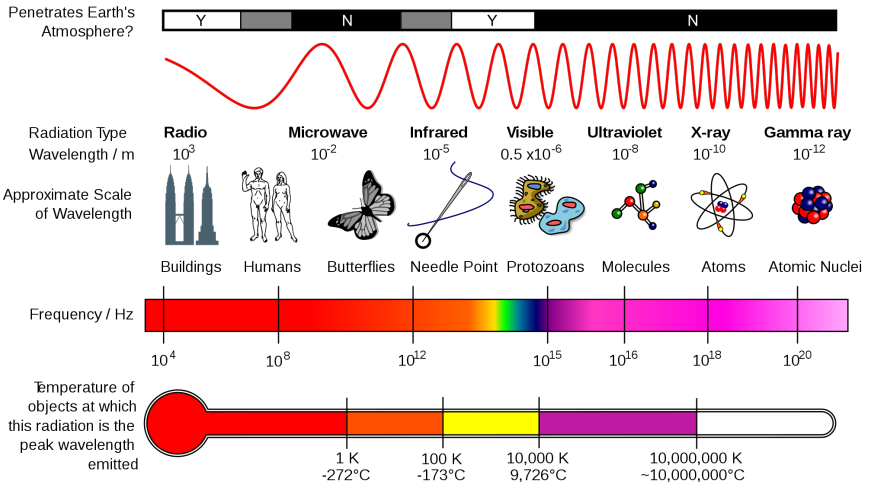


Fig. 2. The spectrum of visible light, its location in the spectrum of electromagnetic waves and the information on which parts of this spectrum reach the earth surface. Source: [18].

In the case of photography or optics, it is usually necessary to calibrate the device to reflect the white correctly, and thus all other colours, based on the ambient light. In modern devices, there are often predefined profiles that offer the selection of temperature values in the average range. However, there may be cases when these values are far from the existing conditions, despite being set well.

In professional photography, the so-called grey cards, referring to surfaces reflecting 18% of the light falling on them, are used. This solution was introduced and widespread by the Kodak company (card R-27 [5]). The choice of the above value is not accidental. Each color can be defined by the parameters determining the electromagnetic wave's physical properties or by a subjective evaluation, a sensory representation related to the organ of vision. Human white light consists of a mixture of wavelengths ranging from 380 to 790 nanometers, and this is related to the spectrum of solar radiation reaching sea level [3]. Different animals see different parts of the light spectrum and can use colour perception systems other than that of humans. In Figure 2 the spectrum of the visible light and its location in the wider spectrum of electromagnetic waves is shown together with the information on which parts of this spectrum reach the earth surface. The spectrum of light emitted from a surface depends on the spectrum of the incident light and on the physical features of the surface itself, which influence the light reflection. The relation of the light wave and the human perception of colour is influenced by the

phenomenon of *colour metamerism*, which consists in that a given colour impression can be received by various combinations of light wave lengths (for example, a yellow colour is perceived when the yellow light is present and also when red and green lights are observed together). Therefore, the perceived colour of the surface can strongly depend on the type of the light source which illuminates it.

In the case of microscopic photography, we deal with relatively homogeneous illumination going from the source to the lens. In addition, the light stream is targeted and often has a bounded region of incidence, which ensures that different microscopes produce similar images. However, this is not always the case, and similarly as for the cameras, it is recommended to calibrate the microscope before each measurement series. Such a process is usually carried out by performing the colour correction for an image of a clean glass and by selecting the appropriate settings based on the known parameters of the optics of the device.

The calibration process ensures that the measurements are comparable to each other over time; however, slight differences between devices introduce some uncertainty concerning the mutual similarity of the results. Thus, a microscopic slide made with a calibrated device from one manufacturer is not identical in colour to a slide made with a device of another one. In addition, the calibration procedure takes time that cannot be omitted in the case of regular measurements of a large number of samples.

The idea behind this research was to bring about a situation in which it is possible to collect images from the microscope without the necessity to carry out the calibration process, and without the necessity to limit the comparison of images to only images from the same device or the same type of the optical acquisition system used. Freeing oneself from these limitations became the basis for developing a solution based precisely on the mechanism of white balance. Issues related to this have already been raised before, inter alia, in [4, 10, 17].

We have founded our study on the *retinex* theory[1] originally introduced in [14,15,19]. This theory gave rise to the White Patch Retinex algorithm for enhancing the colour constancy [14, 19, 22]. It underwent intensive development, see for example [2, 21] in which the *retinex* theory was discussed. In this paper, the *retinex* algorithm has been modified for colour correction of microscopic images.

3.1. Retinex

In the *retinex* theory the human impression of light intensity is treated as depending on the *relative difference* of image brightnesses rather than on the absolute values, which is based on extensive experimental material (described, among others, in [14, 15, 16, 19], and many earlier works cited in [16]). The term *lightness* is used instead of *brightness* or

[1] In the first papers the name of the theory and method, *retinex*, was spelled with lowercase first letter. In later publications the first letter became uppercase, like in the Retinex White Patch algorithm. So, we shall apply the lowercase and uppercase spellings in the respective fragments of the text.

intensity. In the simplest form of the *retinex* algorithm [15] all the random paths leading from a random point in the image to the specified point, in which the lightness value is calculated (the term *lightness* is used in the literature of the *retinex* theory instead of *brightness*). In the first version of the algorithm the relative value of the lightness resulted from the comparison of brightnesses on the individual paths with the value of the lightness of a specified pixel. Considering all the paths is computationally complex but makes it possible to perform a full analysis of an image. The result is the average of the quotients of values of all the subsequent lightness value changes along the paths. It is described by the so calculated lightness value $L(x)$ over all the paths according to the formula (we shall use a clear description of the *retinex* algorithm from [20]):

$$L(x) \quad = \quad \frac{\sum_{k=1}^{N} L(x; y_k)}{N} \tag{1}$$

where: N – number of all the paths, x – starting point of a path, y_x – final pixel of each path, $L(x; y_k)$ – relative value of pixel lightness for a single path:

$$L(x; y_k) \quad = \quad \sum_{t_k=1}^{n_k} \delta \left[\log \frac{I(x_{t_k})}{I(x_{t_{k+1}})} \right] \tag{2}$$

where: n_k – number of pixels in a single path, t_k – subsequent iterated index of pixels in the range, x_{t_k} – lightness in a current pixel, $x_{t_{k+1}}$ – lightness value in the next pixel, δ – threshold of contrast for the given t, where $t \in [0, 1]$:

$$\delta(s) \quad = \quad \begin{cases} s & \text{if } |s| \geq t \\ 0 & \text{if } |s| < t \end{cases} \tag{3}$$

The idea of the algorithm is to find the largest value along the path. In the case of an analysis path by path, the reset system sets to zero the previously found value, if a new value is greater than the one found previously, so the new value becomes the largest one. Additionally, the algorithm performs the task of assuring that its start takes place in the region where the largest lightness value appears. The details of that concept were described in [21].

A relatively important modification of *retinex*, applied in this research, was the algorithm of Single-Scale Retinex (SSR) [12] (submitted in 1995). It is based on the classic choice of a typical local value of lightness with the *nearest neighbours* method (NN). Is is crucial to take into account each of the channels of the colour model in this process. From the point of view of the efficiency and ease of description of the phenomenon, the HSV model should be more suitable for the *retinex* algorithm than the RGB model. This is related above all with the search for the lightness path, which is directly the V component (Value) in the HSV model. For the imaging task and for the analysis of

a typical image the RGB model is used, however, because the lightnesses in the separate hue channels are considered in *retinex*, according to the existence of separate receptors for the long, middle and short wavelengths in the human visual system.

The general form of *retinex* in a given point for one iteration is described with the formula

$$R_i(x,y) \quad = \quad \log(I_i(x,y)) - \log(I_i(x,y) * F(x,y)) \tag{4}$$

where: I_i – input image for one channel of the source ($i-tego$), F – normalised function of the neighbourhood for the pixels belonging to this neighbourhood.

The function $F(x,y)$ proposed by the author of the algorithm is the classic cross averaging method

$$F(x,y) \quad = \quad \frac{C}{x^2+y^2} \tag{5}$$

where C – normalization coefficient.

Alternatively, the Gaussian function is used in SSR:

$$F(x,y) \quad = \quad C * exp^{\frac{-(x^2+y^2)}{2\sigma^2}} \tag{6}$$

where: σ – scale of the filter (deviation). According to the experiments described in [12], $\sigma = 80$ is a good value for calculations. The use of the convolution operation applied before calculating the logarithm in (4) has also been demonstrated.

3.2. White Patch

White Patch is the method based on the *Retinex* theory, which assumes the full use of the possibilities of the active areas of the eye (the rods), which capture the complete information coming from the light falling on them. The brightest point is the one that reflects 100% of the light from the chosen colour [6].

Taking into account that the input image is most often described in the tri-colour component (RGB), the operation should be performed separately for each component. Due to the fact that all the rods are responsible for the white colour, therefore, the obtained range of stimuli is maximised to the entire spectrum by proportionally changing the values. The method is then based on adopting the following transformed data format:

$$R_{\max} = \max_{x,y} R(x,y) \ , \tag{7}$$

$$G_{\max} = \max_{x,y} G(x,y) \ , \tag{8}$$

$$B_{\max} = \max_{x,y} B(x,y) \ . \tag{9}$$

where: (x, y) – coordinates of the point with maximum value lightness of the given colour component, respectively, red (R), green (G) or blue (B).

The White Patch method has been crossed out for digital photography due to the lack of unequivocal effectiveness for both grayscale and colour scale images. However, microscopic images in the processing are ultimately converted to grayscale, so in this case, it is possible to implement a combination of these methods.

The White Patch method for the path mechanism is realised in an the way analogical to the *retinex* algorithm. For the image with high resolution and compression ratio, and with the necessity of representing the hue in wider ranges, the Multi-Scale Retinex (MSR) has been introduced [23]. It is a solution founded on weighted summing of single results for each path separately:

$$R_{\mathrm{MSR}_i} = \sum_{n=1}^{N} \omega_n R_{n_i} \,, \tag{10}$$

$$R_{n_i} = \log(I_i(x, y)) - \log(I_i(x, y) * F(x, y)) \,, \tag{11}$$

where: N – scale of the solution with respect to the coefficient σ (number of SSR components), ω – weight for each scale.

4. Normal Patch Retinex

Taking into account the above information, an algorithm under the name *Normal Patch Retinex* (NPR) is proposed that uses the advantages of the White Patch and Retinex algorithm, based on normalisation using the luminance values of both base algorithms and chrominance matching to obtain an optimised base algorithm. The preparation method is based on the modification of the above-mentioned algorithms.

To understand the procedure, one can visualise the sequential operations according to the scheme shown in Fig. 3.

The source image is read in at the beginning. In the second step, this image is processed with the classic histogram normalization, to find its corrected brightness. In the third step, which can be performed in the parallel way with the second step, the original input image is processed with the White Patch Retinex method. The images from the second and third step are stored separately. These average value of these two images, pixel by pixel, are calculated and stored as the *Average Illuminant*.

The original, source image is then transformed by colour balance adjustment, with the use of the *Average Illuminant* image. The image obtained in this way is stored as the result of the white balance equalization process.

This algorithm was implemented, on the basis of the results received in the trials with the individual algorithms an in the combined methods, presented in Table 2.

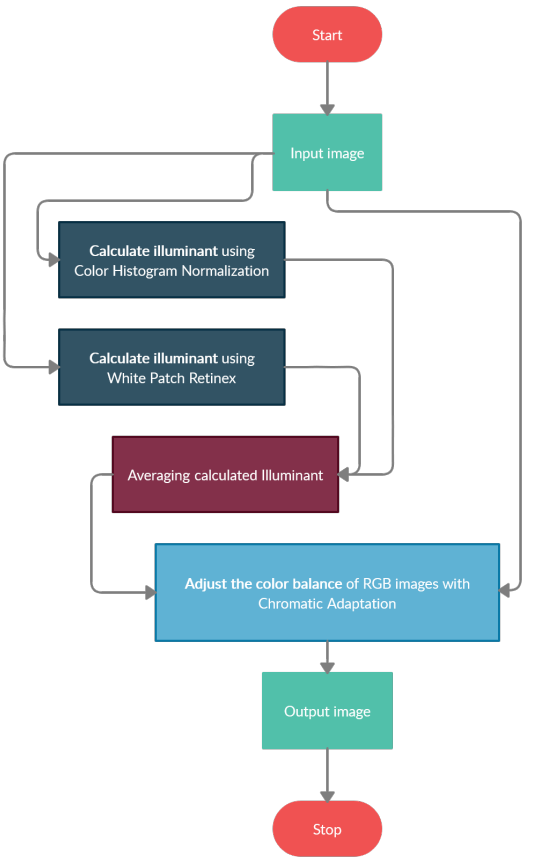


Fig. 3. Diagram representing the general outline of the Normal Patch Retinex algorithm.

The received results indicate univocally that the White Patch Retinex algorithm and the Normal Patch Retinex proposed in this paper yield the most profitable results for the set of the tested images, from the viewpoint of the white balance. The IHC and HPS staining was considered here.

If the whole available set of images and the unification of the method for each type of images acquired with the microscopic method, the proposed Normal Patch Retinex algorithm clearly appears as the most efficient one (Table 3).

Angular error is a helpful metric to evaluate the estimation of an illuminant against the ground truth. The smaller the angle between the illumination determined for the

Tab. 2. Average values of the angular error for the input images of each set for the applied white balance methods.

Method	IHC	HPS
Original	0.48	1.65
Mean Shift Gray Pixel	0.98	4.14
Color Histogram Normalisation	1.92	2.22
Gray World	0.84	5.30
Cheng's Principal Component Analysis	1.73	3.25
White Patch Retinex	**0.47**	1.63
All Gray Pixels	2.71	1.64
YUV Gray Pixels	0.73	3.59
Normal Patch Retinex	1.07	**1.37**

Tab. 3. Standard deviation and averages of the angular error for all input images.

Method	Angular Error	Standard deviation
Original	1.87	1.66
Mean Shift Gray Pixel	2.30	1.93
Color Histogram Normalisation	1.95	0.84
Gray World	2.76	2.09
Cheng's Principal Component Analysis	2.11	1.26
White Patch Retinex	1.86	1.67
All Gray Pixels	2.55	3.60
YUV Gray Pixels	1.67	1.58
Normal Patch Retinex	**1.32**	**0.77**

ground truth and the estimated illumination, then the better the quality of the estimate. To better understand how to use the determined luminance value, make assumptions as illustrated in Fig.4.

5. Results

Images from the slide sets described in Section 2 were used for testing (sample images are shown in Fig. 5). The images were mixed between sets to verify the effectiveness of the adapted algorithm.

In the present study, the original form of staining does not matter much for the algorithms used, due to that the images are transformed to the grayscale.

For research and comparison purposes, the methods used mainly in digital photography to balance white in input images were used here. The summary can be found in

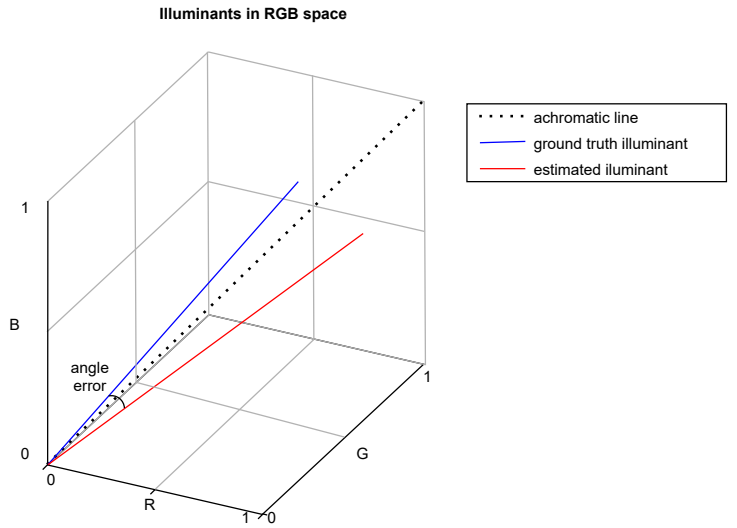


Fig. 4. Luminance lines in the colour space depending on the RGB parameter and the estimated luminance value converted from the reference value.

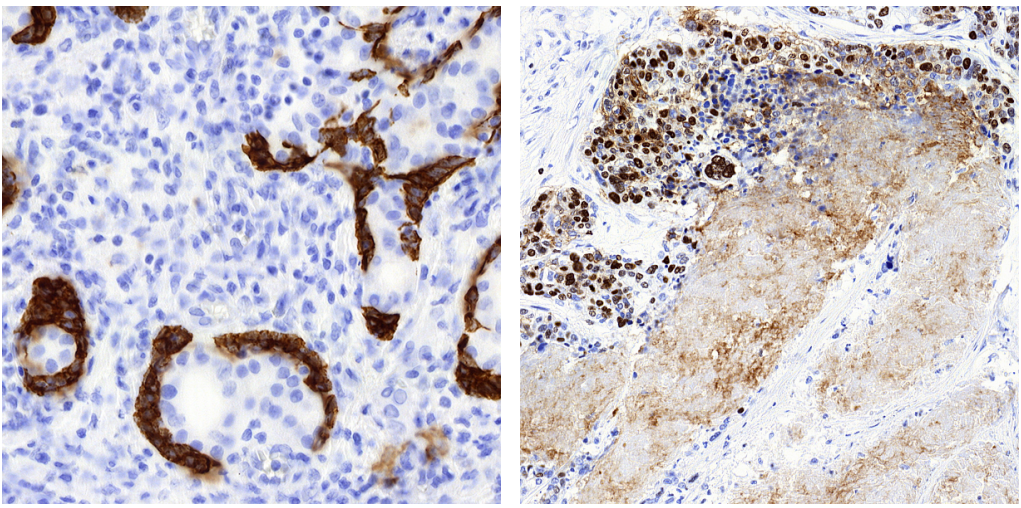
Table 2. As it can be seen, the most appropriate white balance coefficients were achieved for three methods: the aforementioned retinex method, the method based on the White Patch, and the grayscale method for the HSV colour space.

Taking into account the above results, it was assumed that the combination of the mentioned methods might be a solution that makes further use of medical images independent from the parameter of the amount of light or method used for microscopic data acquisition. The first step is to convert the input image to grayscale.

The Table 3 contains the summary results of the comparison of the proposed algorithm in comparison with popular algorithms used in colour photography. For each value of the angular error, standard deviations were calculated for the entire study population consisting of 200 input images. The result achieved by the Normal Patch Retinex algorithm is by far the best with the smallest deviation of the results of the tested samples from the mean value.

In the case of implementing the solution for microscopic images, it was assumed that the converted grayscale image should be normalised. The transfer of RGB colours to the grayscale space ensures that the components retain their values despite the expansion of the colour spectrum.

In traditional photography, algorithms search for the darkest and brightest places to



a **b**

Fig. 5. Comparison of two slides made under different laboratory conditions with different staining methods. IHC samples, stained with: (**a**) CK34, (**b**) KI67.

determine the range of values across the entire set of spaces. To extend the histogram for a grey image in a similar way, it is necessary to find pixels with dark and light values, respectively. The difference of this algorithm for microscope slides is that the light surfaces are the passage of light through the object, and the dark ones – places where the beam was stopped.

In other words, while in photography, the light falling on a bright object reflects off it and hits the sensor in the lens, in the case of a microscopic object, the reflected light does not reach the lens because it reflects off the glass. The algorithm should, therefore perform the colour assignment in the opposite way than the value of the scanned slide indicates.

Referring to the microscopic output image constructed in this way, the result is an image with the most optimally matched colour balance. However, to verify the thesis, the Normal Patch Retinex needs to be implemented and carried out for the collected data set. Fig. 6 shows sequentially numbered results of the algorithm's operation.

According to the aforesaid method of calculating the luminance in the RGB space, the correct determination of the estimated value requires to average the value calculated by the algorithm and the reference luminance value. The difference between the offset angles of both luminance is the value sought for which the white balance method using chrominance adaptation can be later used.

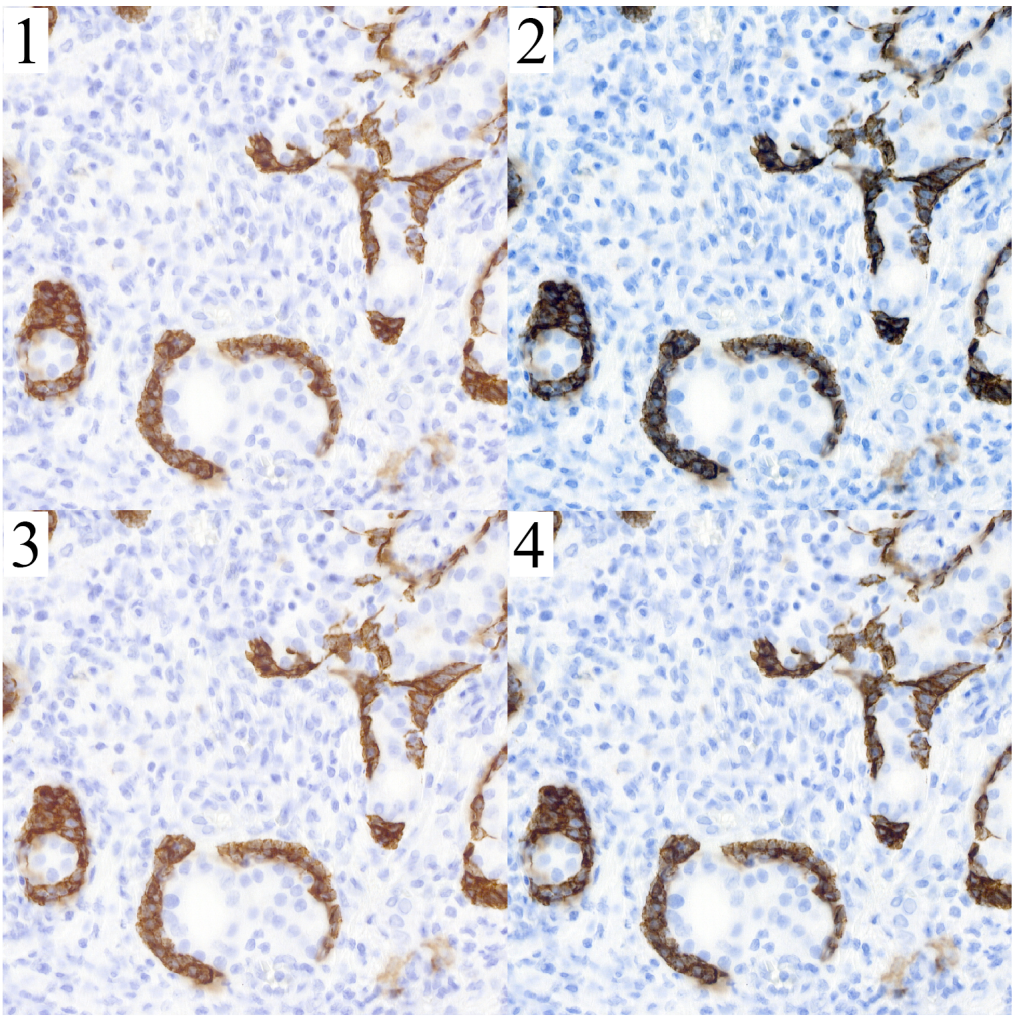


Fig. 6. Comparison of the effects of white correction on the basis of a preparation with immunohistochemistry staining. Algorithm numbering: (**1**) original image, (**2**) white balancing with Colour Histogram Normalisation, (**3**) white balancing with White Patch Retinex, (**4**) white balancing with Normal Patch Retinex.

6. Conclusion

The proposed white balance correction algorithm in microscopic images allows for quick and effective colour equalisation. This algorithm does not require the prior preparation of input data or other pre-processing methods. The big advantage of the Normal Patch Retinex algorithm is its speed, full automaticity and ease of use.

The presented algorithm solves the problem of white balance equalization in a way dedicated to microscopic imaging. Previously, the algorithms used in colour photography were used for medical imaging. This algorithm properly corrects the white balance in images of tissues stained with different methods. The algorithm can be successfully used in the process of pre-treatment of single scans of microscopic slides or the entire series of microscopic images. The use of NPR to align the colour space of a series of images allows obtaining a consistent colour space for all processed images.

References

[1] P. Baldevbhai. Color Image Segmentation for Medical Images using L*a*b* Color Space. *IOSR Journal of Electronics and Communication Engineering*, 1(2):24–45, 2012. doi:10.9790/2834-0122445.

[2] M. Bertalmío, V. Caselles, and E. Provenzi. Issues about Retinex theory and contrast enhancement. *International Journal of Computer Vision*, 83(1), Jun 2009. doi:10.1007/s11263-009-0221-5.

[3] P. Biecek. Perception of colours (in Polish). In *Odkrywać! Ujawniać! Objaśniać! Zbiór esejów o sztuce prezentowania danych*, pages 67–84. Fundacja Naukowa SmarterPoland.pl, 2016. `http://biecek.pl/Eseje/indexKolory.html`.

[4] R. Davis. A correlated color temperature for illuminants. *Bureau of Standards Journal of Research*, 7(4):659, 1931. doi:10.6028/jres.007.039.

[5] Eastman Kodak Company. KODAK Gray Card / R-27, 2020. `https://www.kodak.com/en/motion/page/gray-cards`. [Accessed Oct 2020].

[6] H. Garud, A. Ray, M. Mahadevappa, et al. A fast auto white balance scheme for digital pathology. *2014 IEEE-EMBS International Conference on Biomedical and Health Informatics, BHI 2014*, pages 153–156, 2014. doi:10.1109/BHI.2014.6864327.

[7] B. Gilbert et al. MIRAX (MRXS). In Goode et al. [9]. [Accessed Oct 2020]. `http://openslide.cs.cmu.edu/download/openslide-testdata/Mirax/`.

[8] A. Goode, B. Gilbert, J. Harkes, et al. OpenSlide: A vendor-neutral software foundation for digital pathology. *Journal of Pathology Informatics*, 4(1):27, 2013. doi:10.4103/2153-3539.119005.

[9] A. Goode, B. Gilbert, J. Harkes, et al., editors. *OpenSlide*, 2020. [Accessed Oct 2020]. `https://openslide.org/`.

[10] K. Hasna Panikkaveettil and M. V. Beena. A survey on color normalization approach to histopathology images. *International Journal of Advanced Engineering Research and Science*, 3(4):103–105, 2016. `https://www.neliti.com/publications/258867/`.

[11] M. Holek. File:Color temperature.svg. In *Temperatura barwowa – Wikipedia, wolna encyklopedia* [26]. From Wikimedia Commons. License: creative commons cc-by-sa 2.5 Poland [Accessed Oct 2020]. `https://commons.wikimedia.org/wiki/File:Color_temperature.svg`.

[12] D. J. Jobson, Z. Rahman, and G. A. Woodell. Properties and performance of a center/surround retinex. *IEEE Transactions on Image Processing*, 6(3):451–462, 1997. doi:10.1109/83.557356.

[13] E. Y. Lam and G. S. K. Fung. Automatic white balancing in digital photography. In Lukac R., editor, *Single-Sensor Imaging: Methods and Applications for Digital Cameras*, pages 267–294. CRC Press, Boca Raton, 2017. doi:10.1201/9781315219363.

[14] E. Land. The retinex theory of color vision. *Scientific American*, 237(6):108–128, Dec 1977. doi:10.1038/scientificamerican1277-108.

[15] E. Land and J. McCann. Lightness and retinex theory. *Journal of the Optical Society of America*, 61(1):1–11, Jan 1971. doi:10.1364/JOSA.61.000001.

[16] E. H. Land. The Retinex. *American Scientist*, 52(2):247–264. `http://www.jstor.org/stable/27838994`.

[17] M. Macenko, M. Niethammer, and J.and others Marron. A method for normalizing histology slides for quantitative analysis. *Proceedings - 2009 IEEE International Symposium on Biomedical Imaging: From Nano to Macro, ISBI 2009*, pages 1107–1110, 2009. doi:10.1109/ISBI.2009.5193250.

[18] Inductiveload (Wikimedia user). A diagram of the EM spectrum, showing the type, wavelength (with examples), frequency, the black body emission temperature. adapted from `em_spectrum3-new.jpg`, which is a NASA image, Oct 2007. `https://commons.wikimedia.org/wiki/File:EM_Spectrum_Properties.svg`. From Wikimedia Commons. License: Attribution-ShareAlike 3.0 Unported cc-by-sa 3.0 [Accessed Jul 2020].

[19] J. McCann. Color sensations and color perceptions. In *Proc. 24th Asilomar Conf. Signals, Systems and Computers ACSSC*, volume 1, pages 408–412, Pacific Grove, CA, USA, 5-7 Nov 1990. IEEE Computer Society. doi:10.1109/ACSSC.1990.523369.

[20] A. B. Petro, C. Sbert, and J.-M. Morel. Multiscale retinex. *Image Processing On Line*, pages 71–88, 2014. doi:10.5201/ipol.2014.107.

[21] E. Provenzi, L. De Carli, A. Rizzi, and D. Marini. Mathematical definition and analysis of the Retinex algorithm. *Journal of the Optical Society of America A*, 22(12):2613–2621, Dec 2005. doi:10.1364/JOSAA.22.002613.

[22] E. Provenzi, C. Gatta, M. Fierro, and A. Rizzi. A spatially variant White-Patch and Gray-World method for color image enhancement driven by local contrast. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 30(10):1757–1770, 2008. doi:10.1109/TPAMI.2007.70827.

[23] Z. Rahman, D. J. Jobson, and G. A. Woodell. Multi-scale retinex for color image enhancement. In *Proc. 3rd IEEE Int. Conf. Image Processing*, volume 3, pages 1003–1006, Sep 19, 1996. doi:10.1109/ICIP.1996.560995.

[24] M. Saha, S. Agarwal, I. Arun, et al. Histogram based thresholding for automated nucleus segmentation using breast imprint cytology. In S. Gupta, S. Bag, K. Ganguly, et al., editors, *Proc. 1st Int. Conf. Advancements of Medical Electronics ICAME 2015*, Lecture Notes in Bioengineering, pages 49–57. Springer, Hamburg, Germany, 29-30 Jan 2015. doi:10.1007/978-81-322-2256-9_5.

[25] J. Thiran and B. Macq. Morphological feature extraction for the classification of digital images of cancerous tissues. *IEEE Transactions on Biomedical Engineering*, 43(10):1011–1020, 1996. doi:10.1109/10.536902.

[26] Wikipedia contributors. Temperatura barwowa – Wikipedia, wolna encyklopedia, Nov 2019. `https://pl.wikipedia.org/wiki/Temperatura_barwowa`. [Accessed Oct 2020].

[27] Wikipedia contributors. Color temperature – Wikipedia, The Free Encyclopedia, Dec 2020. `https://en.wikipedia.org/wiki/Color_temperature`. [Accessed Oct 2020].