Instituto Tecnológico de Costa Rica

Vicerrectoría de Investigación y Extensión

Dirección de Proyectos

Informe Final de Proyecto de Investigación

Documento I

Periodo 2010–2011

Análisis por computador de imágenes de geles de electroforesis: métodos avanzados de manejo de meta-información y procesamiento digital de imágenes

5402 1360 2601

Adscrito a:
Escuela de Ingeniería Electrónica

Investigador principal:
Dr. José Pablo Alvarado, EIE

Investigadores:

M.Sc. Alicia Salazar, CIC	Dr. José Enrique Araya, CIC	M.Sc. Johnny Peraza, CIB
Ing. Fabiana Rojas, CIIBI		Dr. Olman Murillo, CIIBI

Estudiantes tesistas:

Pedro Alpízar Salas	Bryant Álvarez Canales	Pablo Barrantes Chaves
Randall Esquivel Alvarado	Edison Fernández Alvarado	David Soto Vásquez

4 de julio, 2012

Agradecimientos

Este proyecto ha contado con la valiosa colaboración de asistentes financiados por medio de las becas asistente especial. En especial queremos agradecer a los estudiantes de ingeniería electrónica Pablo Barrantes, David Soto, Antonio Aguilar, Bryant Álvarez, Edison Fernández, Pedro Alpízar, Randall Esquivel, Mauricio Caamaño, Bayron Monge, Eduardo Corrales, Diego Sánchez y Andrés González; los estudiantes de ingeniería en computadores Javier Montoya y Edward Jiménez, así como a las estudiantes de ingeniería en biotecnología Shirleny Sandoval, Rossi Guillén y Milenna González. Sin el arduo trabajo de estos estudiantes, no se hubiese podido avanzar en el proyecto.

Culminaron sus tesis de licenciatura en el contexto del proyecto José Antonio Aguilar, Pablo Barrantes, David Soto, Bryant Álvarez Edison Fernández, Pedro Alpizar y Randall Esquivel, todos ellos asesorados por el investigador principal del proyecto. Debe destacarse que el trabajo de Randall Esquivel fue reconocido con el Premio Asoelectrónica 2011 a los mejores Proyectos de Graduación de Ingeniería Electrónica. Todos ellos se han reconocido como co-autores de este informe, pues métodos y resultados de sus tesis se han reproducido en este informe.

Resumen

La caracterización molecular se entiende como el proceso mediante el cual se determinan atributos conspicuos de una molécula en particular. Este trabajo continua con el desarrollo de algoritmos a ser utilizados en un sistema computacional para apoyar labores de caracterización molecular, que permite asociar características particulares a organismos de acuerdo a la composición de moléculas como el ADN, ARN, proteínas, etc. Dicha composición se evalúa en este caso particular a través de los llamados geles de electroforesis, cuyas imágenes presentan usualmente bajos contrastes y distorsiones que dificultan su uso.

Este proyecto aporta nuevos algoritmos y estructuras para la administración de información a ser integrados en una herramienta informática que persigue potenciar la utilidad de las imágenes de geles de electroforesis, reduciendo por un lado los tiempos invertidos en su análisis y por otro lado incrementando la confiabilidad y robustez de los resultados obtenibles, compensando errores originados por el factor humano, y facilitando el rescate de información oculta. Además, la herramienta se encarga de manejar metainformación asociada a las imágenes, que incluye todo el proceso desde la toma de la muestra hasta la captura misma de las imágenes, con el fin de permitir posteriormente la búsqueda de datos por medio de esa metainformación.

El proyecto culmina luego de 5 años, que es aproximadamente una mitad del tiempo en que proyectos similares en Irlanda y Austria han estado publicando sus avances. Resta por iniciar el proceso de desarrollo de producto, que escapa al marco de trabajo del proyecto de investigación. Este informe muestra los aportes alcanzados en el proyecto en cuanto a algoritmos y métodos, capaces de brindar mayor automatización en laboratorios de biología molecular.

Los resultados principales se concentran, por un lado, en la consolidación de la arquitectura propuesta anteriormente para todo el sistema, la puesta en marcha de un sistema adaptativo de captura, compensación de distorsiones ópticas y de perspectiva en la detección de carriles, un sistema de difusión con mejoramiento de coherencia con filtro orientado para corregir la distorsión del efecto sonrisa, y un complejo sistema de detección de las bandas en un carril que combina análisis en el espacio de escala con procesos de optimización. Por otro lado, se diseñó el sistema de bases de datos distribuidas y métodos avanzados de consulta basados en minería de datos. Finalmente, se establecieron los protocolos para generación de electroforesis en gel de gradiente desnaturalizante, que se utilizaron en el

análisis de diversidad bacteriana en muestras de suelo y agua.

Palabras clave: *análisis de imágenes, segmentación de carriles, eliminación de fondo, efecto sonrisa, band detection, PCR, electroforesis, caracterización molecular de organismos, bases de datos, DGGE*

Abstract

Molecular characterization is the process in which conspicuous attributes of specific molecules are determined. This work continues with the development of algorithms to be used in a computational system to support tasks of molecular characterization, which allows to link particular features to organism according to the composition of molecules such as DNA, RNA, proteins, etc. Such composition is evaluated in this particular case through electrophoresis gels, whose images usually exhibit low contrast and distortions that draw their use difficult.

This project contributes with new algorithms and structures for information management to a computational tool that improves the electrophoresis gel images utility, by reducing the invested times in their analysis and by augmenting the reliability and robustness of the obtainable results,, compensating errors originated in human factors, and making easier the recovery of hidden information. Additionally, the tool is capable of managing metainformation associated to the images, including the whole process from the sample taking to the image capture, with the goal to allow advance search of data based on that metainformation.

This project finishes after five yearis now three years young, which is approximately one half of the time of existence of similar projects in Ireland and Austria that have been publishing their results. This report shows valuable contributions in the proposed algorithms, capable to improve the degree of process automation in the molecular biology laboratories

The main results reside, on one side, in the consolidation of the proposed architecture for the whole system, the creation of an apative capture system, the compensation of optical and perpective distortions in the lane detection, the coherence enhanced diffusion system improved by orientation filtering to correct the distortions caused by the smile-effect, and a novel system for the band detection in a lane which combines scale-space analysis with optimization processes. On the other side, a distributed data base system has been designed and advanced query methods based on data mining are proposed. Last, but not least, protocols have been established for the generation of denaturing gradient gel electrphoresis, which have been used in the analysis of bacterial diversity in soil and water samples.

Keywords: *image analysis, lane segmentation, background elimination, smile effect, band detection, PCR, electrophoresis, molecular characterization, data bases, DGGE*

Índice general

Índice de figuras

Índice de tablas

Índice de abreviaturas

AFLP Amplified Fragment Length Polymorphism

CED Coherence Enhanced Diffusion

DGGE Denaturing Gradient Gel Electrophoresis

PCA Principal Component Analysis

PCR Polymerase Chain Reaction

Capítulo 1

Introducción

El verbo *caracterizar*, según el Diccionario de la Real Academia de la Lengua Española, significa en su primera acepción "determinar los atributos peculiares de alguien o de algo, de modo que claramente se distinga de los demás". La caracterización *molecular* se entiende entonces como el proceso mediante el cual se determinan atributos conspicuos de una molécula en particular. Este trabajo se concentra en el diseño de algoritmos a ser utilizados en un sistema computacional para apoyar labores de caracterización molecular de organismos, que, extendiendo las ideas anteriores, permite asociar características particulares a organismos de acuerdo a la composición de moléculas como el ADN, ARN, proteínas, etc.

Para el caso particular de caracterización del ADN, el método más preciso es la *secuenciación genética*, que extrae la secuencia de bases nitrogenadas constituyentes de esta molécula: Adenina, Guanina, Citosina y Timina. El análisis de las complejas cadenas obtenidas de la secuenciación, representadas con series de los caracteres A,G,C y T ha dado origen al área de la bioinformática, donde se buscan algoritmos eficientes que permitan, entre otros, almacenar, reconocer y gestionar dichas secuencias y sus patrones. Si bien es cierto la precisión de la secuenciación es elevada, también lo son los costos asociados en equipo, software, tiempo, reactivos y químicos requeridos para el proceso completo de caracterización. Dicha precisión no siempre es necesaria en las aplicaciones encontradas día a día en los laboratorios de biología molecular.

El proceso de desarrollo de los algoritmos para este proyecto ha tomado como referencia el trabajo de biólogos moleculares y genetistas en

1. la tipificación de cepas bacterianas
2. el análisis y caracterización de metagenomas bacterianos presentes en suelos
3. la verificación de pureza entre individuos de plantaciones forestales

En estos proyectos los métodos más utilizados para la caracterización molecular se basan en el análisis de imágenes de geles de electroforesis. La electroforesis denota procesos de separación molecular basados en características como dimensión, carga eléctrica, forma, etc. de las moléculas [12]. El gel es un polímero entrelazado de porosidad controlable,

usualmente agarosa para moleculas con cientos de pares de bases, o poliacrilamida para moléculas más pequeñas, a través del cual un potencial eléctrico fuerza el desplazamiento de las moléculas a diferentes velocidades dependiendo de su carga eléctrica.

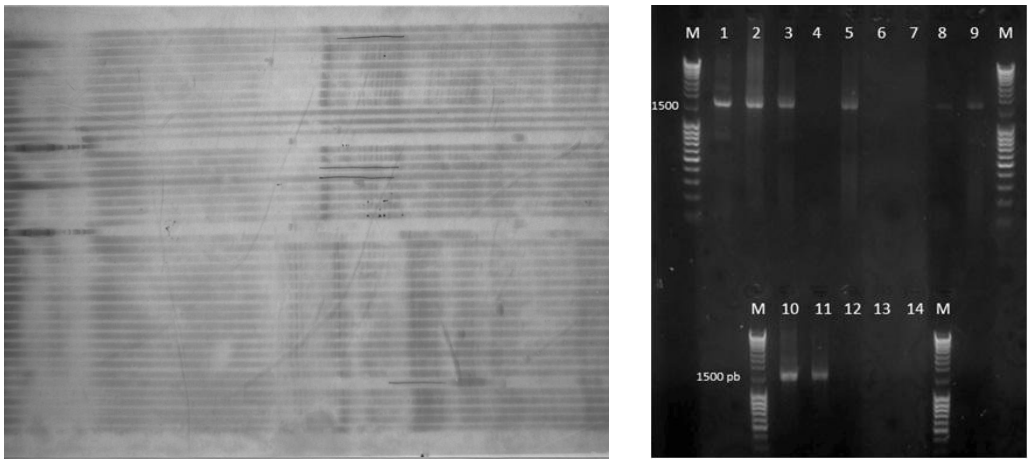


Figura 1.1: Geles de electroforesis. A la izquierda se ilustra un gel utilizando la técnica AFLP. A la derecha una prueba de amplificación de ADN.

La figura 1.1 muestra dos ejemplos de los geles generados durante el proyecto. En éstos llamados geles unidimensionales a cada muestra de ADN corresponde un carril, calle o pista (*lane* o *track* en inglés). Sobre el carril se conforman bandas, cuya distribución depende de la metodología de análisis empleada y de la molécula particular de ADN bajo escrutinio, de modo que el patrón de bandas es característico del organismo evaluado. En todos los métodos, cada banda corresponde a la presencia de moléculas con una longitud de pares de bases específica.

1.1 Antecedentes

El análisis manual de las imágenes de geles de electroforesis es una tarea extenuante para el sistema visual humano, por tratarse de imágenes monocromáticas de bajo contraste, en las que la decisión sobre la presencia o ausencia de bandas requiere de gran concentración y esfuerzo visual. Por otro lado, la determinación de la presencia o ausencia de bandas, así como la correspondencia de bandas entre carriles, es una tarea propensa a errores por los problemas de distorsión de las imágenes, y efectos como la fatiga del observador.

El presente proyecto constituye la continuación de la actividad de fortalecimiento titulada “Análisis automatizado de patrones de ADN para la caracterización molecular” [6], y del proyecto “Análisis por computador de imágenes de geles de electroforesis para la caracterización molecular de organismos” [7]. En ambos se hacen aportes a la creación de una herramienta informática que potencie la utilidad de las imágenes de geles, reduciendo por un lado los tiempos invertidos en su análisis y por otro lado incrementando la confiabilidad y robustez de los resultados obtenibles, compensando errores originados en el factor humano, y facilitando el rescate de información oculta.

En esta área, la investigación a nivel mundial se ha concentrado en los problemas de detección de carriles, su alineamiento, corrección del efecto sonrisa, reducción del "fondo", reducción de ruido y normalización, para permitir comparaciones inter-geles, y almacenar en la base de datos la información de modo recuperable por minería de datos.

En los últimos años han salido al mercado productos principalmente orientados a resolver los módulos de análisis de imágenes. Por ejemplo, el LabImage 1D de la empresa alemana Kapelan GmbH provee una colección de módulos de análisis de la imagen únicamente. El Gel-Quant de la empresa australiana Ampl Software provee algunas herramientas básicas de rectificación, detección de carriles y bandas y el cálculo de peso molecular, áreas, masas etc. La empresa belga Applied Math provee un sistema modular, incluyendo al GelCompar II y GelNumerics que en conjunto ofrecen un sistema similar al planteado en este proyecto: no solo el análisis de las imágenes sino un sistema de base de datos para manejar la información generada son necesarios para el uso efectivo de la información. También han salido al mercado productos para la adquisición de las imágenes de geles, como las plataformas MiniBIS de la empresa israelí DNR Bio-Imaging systems.

La existencia de estos productos pone en evidencia la importancia de sistemas de análisis y manejo de información en los laboratorios de biología molecular. Una fortaleza del sistema que se propone en éste documento, que comparte con los productos de Applied Math, es el manejo de la información obtenida en una base de datos. Una ventaja adicional que caracteriza al presente proyecto es que incluye además toda la información del proceso que antecede a los geles en sí. Esto abre espacios de investigación en procesos de minería de datos basados en la meta-información, combinados con aquellos basados propiamente en la información de los geles. La utilización de herramientas de software libre en la implementación permite además bajar costos de licencias de desarrollo y de operación. Por otro lado, la utilización de una interfaz web como elemento central de la aplicación es otra ventaja de la presente propuesta, puesto que permite implementar todo el sistema en una plataforma específica, y ser utilizada desde un navegador desde cualquier otra plataforma.

1.2 Prototipo computacional para el tratamiento de imágenes de geles

La figura 1.2 ilustra la separación conceptual del proyecto en dos bloques: el primero relacionado con la gestión de la información y el segundo encargado del análisis de las imágenes digitales. La investigación ha estado orientada por el desarrollo de un prototipo que permita:

1. Mejorar las imágenes de geles de electroforesis adquiridas de forma digital, de modo que un especialista en su análisis pueda extraer de ellas la información relevante con mayor facilidad. Esto implica el diseño de procesos de mejoramiento de contraste y de rectificación geométrica, entre otros.

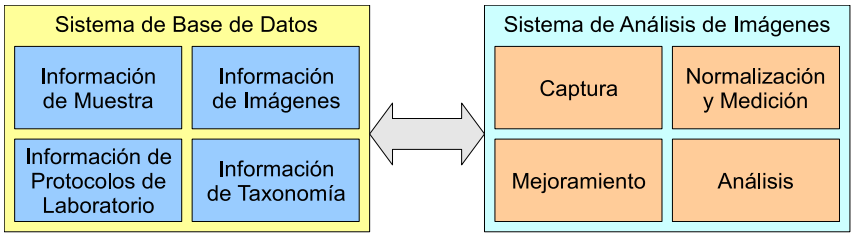


Figura 1.2: Bloques involucrados en el proyecto

2. Analizar automáticamente las imágenes mejoradas utilizando técnicas de visión por computador (es decir, de procesamiento y análisis digital de imágenes, y de reconocimiento de patrones). El diseño e identificación de dichas técnicas y sus fundamentos teóricos deben considerar en todo momento las necesidades reales de los expertos en biología molecular que las utilicen, de modo que el sistema entregue información útil utilizando formatos adecuados.

3. Almacenar y proveer métodos de acceso a toda la información y datos involucrados en los procesos de captura de imágenes y los resultados de su análisis, por medio de un sistema de base de datos. Este sistema debe considerar todos los aspectos sobre protección de la propiedad intelectual necesarios para que los investigadores puedan depositar allí sus datos con seguridad de que no sean invisibles a terceros, y poder dar acceso condicionado a terceros si se desea. Por otro lado, debe considerar que parte de los datos a manejar son imágenes, y por tanto con requerimientos de espacio de almacenamiento y de ancho de banda de transmisión especiales. Además, el sistema debe almacenar y administrar la meta-información de las imágenes obtenidas, como por ejemplo, el método particular utilizado para su generación, el laboratorio e investigador a cargo, fecha, etc., que permitan flexibilizar los mecanismos disponibles de búsqueda y acceso.

El módulo de base de datos debe almacenar toda la información obtenida en el proceso de obtención de los geles. Este proceso incluye

1. los datos del origen del muestra (provincia, cantón, individuo, coordenadas GPS, etc.),
2. los protocolos utilizados para extraer dicha muestra, almacenados con control de versiones,
3. identificadores para las muestras extraídas,
4. los protocolos de extracción de ADN de las muestras anteriores, con datos de valoración del proceso de extracción, incluyendo imágenes de geles de electroforesis,
5. los protocolos utilizados para el análisis molecular de las muestras, especificando las técnicas concretas y los parámetros utilizados
6. los protocolos del corrimiento de los geles
7. los protocolos de captura de las imágenes, incluyendo tipo de cámaras y sus parámetros.
8. las imágenes de geles de electroforesis como tales.

Los componentes del módulo de procesamiento de imágenes se ilustran en la figura 1.3. Esta estructura fue propuesta en los proyectos antecesores [6, 7]. Los primeros tres bloques:

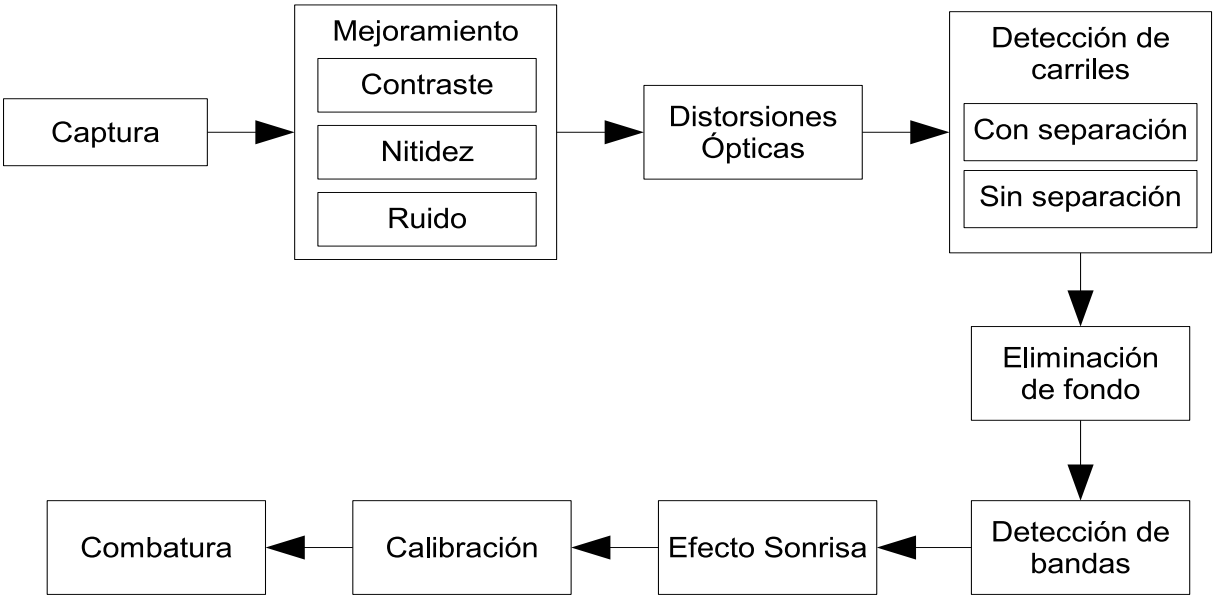


Figura 1.3: Bloques principales del módulo de procesamiento de imágenes.

captura, mejoramiento, y eliminación de distorsiones ópticas, son problemas genéricos y por tanto se encuentra suficiente literatura al respecto en variedad de contextos. Los siguientes bloques tratan problemas particulares del análisis de geles de electroforesis unidimensionales, y se profundizará en ellos en los siguientes capítulos.

La enorme variabilidad de técnicas existentes para realizar la caracterización molecular con geles de electroforesis (AFLP, REP-PCR, PCR, Micro-Satélites, PAGE, DGGE, etc.), así como las distintas técnicas de tinción y revelado, produce imágenes cada una con características particulares que dificultan el uso de técnicas de procesamiento digital universales. De hecho, cada uno de los artículos científicos encontrados se concentran en un único tipo particular de imágenes en una aplicación, lo que dificulta la comparación objetiva entre los distintos algoritmos. En este proyecto en particular, se tratan imágenes de AFLP (con y sin separación entre carriles) pues estas incluyen los mayores grados de dificultad por el número de bandas y la falta de contraste. La adaptación de los algoritmos a los otros métodos será entonces directa.

La figura 1.4 muestra una captura de pantalla de la aplicación desarrollada. El sistema se ejecuta en un servidor remoto y brinda acceso a la información a través de internet.

1.3 Objetivos y estructura del informe

El objetivo general del presente proyecto ha sido incorporar a la herramienta desarrollada en los proyectos antecesores opciones avanzadas para el manejo de meta-información y para el análisis de imágenes de geles de electroforesis utilizadas en la caracterización

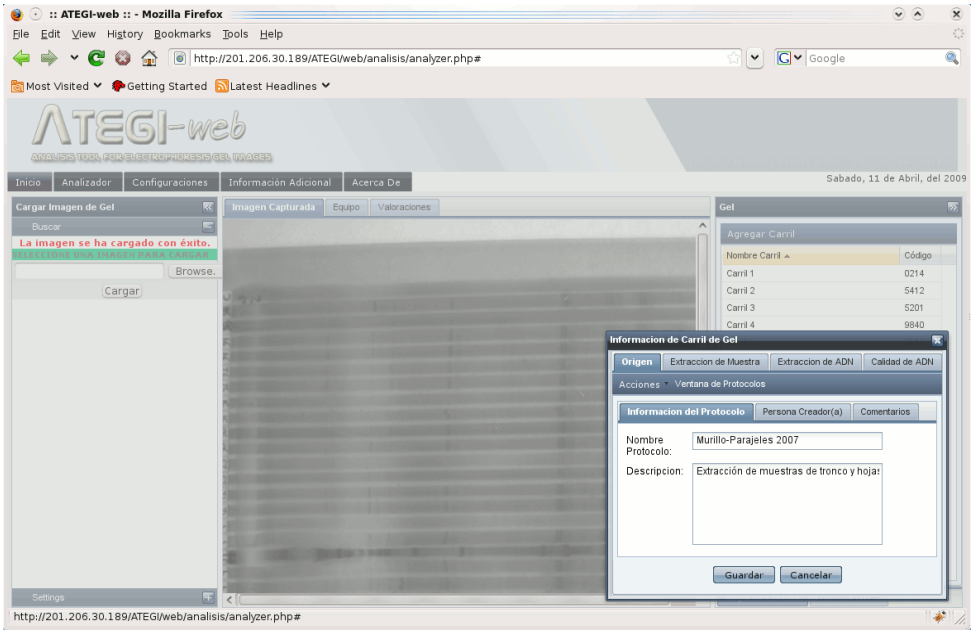


Figura 1.4: Captura de pantalla del software ATEGI desarrollado en este proyecto y sus antecesores.

molecular de organismos, de modo que se mejore la usabilidad del sistema en laboratorios de biología molecular.

Los objetivos específicos que condujeron la investigación se enumeran a continuación:

1. Mejorar la seguridad en el manejo de los datos del sistema por medio del concepto de "administración de roles".
2. Extender la arquitectura del sistema para manejar bases de datos distribuidas.
3. Agregar a la interfaz de usuario modos adicionales para la manipulación de las imágenes, la interacción con el sistema y la presentación de la información almacenada y generada.
4. Identificar automáticamente los carriles de control con base en datos almacenados.
5. Crear un módulo integrado de rectificación y normalización de imágenes de geles.
6. Integrar información contextual en los algoritmos de análisis de imágenes de geles.
7. Integrar la meta-información, imágenes y métodos de análisis de diversidad genética bacteriana en muestras ambientales utilizando la DGGE.

En la siguiente sección se presentarán los conceptos teóricos así como la revisión de literatura reciente relacionada con el estado del arte de los métodos utilizados en el proyecto. Seguidamente se presentarán los algoritmos y métodos desarrollados para el proyecto. El capítulo 4 presenta los resultados y análisis principales de los algoritmos propuestos. Debe hacerse énfasis de que el detalle de los métodos y revisión del estado del arte completo escapa al marco espacial de este informe, por lo que se hace referencia constante a las tesis desarrolladas en el contexto del proyecto. Las conclusiones y recomendaciones se resumen en el capítulo 5. Los principales aportes del proyecto se rescatan en el capítulo 6.

Capítulo 2

Marco teórico y estado del arte

Se revisan en este capítulo principios teóricos que fundamentan los diseños planteados o son el punto de partida para las nuevas propuestas desarrolladas como parte de la investigación, y detalladas en el siguiente capítulo.

Para iniciar, en la sección 2.1 se repasan los principios que rigen la formación de las imágenes de electroforesis en gel (o geles de electroforesis). Posteriormente, las secciones 2.2–2.6 cubren los fundamentos teóricos y estado del arte utilizado para desarrollar los módulos indicados en la figura 1.3 relacionados con el procesamiento y análisis de las imágenes digitales. En particular, la sección 2.2 presenta aspectos relacionados con el módulo de captura; la sección 2.3 presenta la herramientas utilizadas en los módulos de compensación de distorsiones ópticas y de detección de carriles; la sección 2.4 está relacionada con la detección del efecto sonrisa y la sección 2.5 con los fundamentos para comprender las propuestas de detección de las bandas.

Se presenta en la sección 2.6 una de las estrategias de optimización y evaluación utilizada en el proyecto, basada en algoritmos genéticos para optimización multiobjetivo.

Finalmente, en la secciones 2.7 y 2.8 se abortan los conceptos asociados a bases de datos distribuidas y minería de datos, utilizados en el proyecto.

2.1 Electroforesis en gel

Los geles de electroforesis son una herramienta de análisis cuyo principio básico es el movimiento controlado de partículas cargadas bajo a acción de un campo eléctrico [15]. Las moléculas en los ácidos nucleicos por lo general poseen carga negativa, por lo que al aplicar una diferencia de potencial entre los extremos del gel de poliacrilamida las moléculas tratarán de avanzar a través del gel desde el cátodo hasta el ánodo. La velocidad de migración de una molécula v es directamente proporcional al campo eléctrico E, a la

carga neta de la molécula z e inversamente proporcional al coeficiente de fricción de f:

$$v = \frac{Ez}{f} \tag{2.1}$$

El coeficiente de fricción f es directamente proporcional a la masa y el tamaño de la molécula, y a la viscosidad del medio, en este caso el gel.

El proceso de electroforesis es realizado generalmente en geles debido a que estos sirven como una rejilla que facilita la separación de las moléculas. Las moléculas que son pequeñas en comparación con los poros del gel se mueven fácilmente a través de éste, mientras que las moléculas grandes tienen su movimiento limitado. La electroforesis se lleva a cabo en placas verticales de gel de agarosa o de poliacrilamida, donde se colocan las moléculas a estudiar en la parte superior, de forma que la dirección de movimiento sea de arriba hacia abajo. Las moléculas pequeñas se mueven rápidamente a través del gel, mientras que las moléculas grandes quedan, en general, cerca del punto de aplicación de la muestra.

En una placa de gel de electroforesis se aplica una serie de muestras que se analizan al mismo tiempo. Los patrones verticales que se observan en la imagen del gel corresponden a los carriles, donde cada uno de estos representa la distribución de moléculas para una muestra en análisis. Las formaciones horizontales oscuras dentro de un carril reciben el nombre de bandas y corresponden a una acumulación de moléculas en una posición particular.

Uno de los usos de los geles de electroforesis es el análisis de compuestos de proteínas, y permite realizar un análisis cualitativo en una imagen por la forma en la que se desplazan las moléculas a través del gel [15].

La ubicación de las bandas dentro del carril se utiliza como una forma de comparar la composición molecular del ADN de dos organismos diferentes para determinar qué tan similares son genéticamente.

2.2 Mejora de calidad de imagen desde la captura

Para analizar las imágenes digitales de geles de electroforesis, así como para cuantificar la calidad de cada imagen capturada se utilizan los atributos:

- Intensidad
- Contraste
- Ruido

donde cada una de ellos medir de manera local u holística en una imagen digital:

Local es el valor analizado de un pixel con su respectivo vecindario; excluye todos los píxeles que no forman parte de este conjunto.

Global es el valor analizado para la imagen digital como un todo.

2.2.1 Intensidad

En [2] se define la intensidad como la energía proyectada en un cierto punto (i, j) o en toda la imagen digital I por el sistema de adquisición. Matemáticamente ésto se expresa como

$$L(i, j) = l(i, j)r(i, j) \tag{2.2}$$

donde $L(i, j)$ es la intensidad de un píxel de la imagen en las coordenadas (i, j) (o en toda la imagen digital) y en teoría puede tomar valores en el rango $0 < L(i, j) < \infty$. La función $l(i, j)$ es la cantidad de energía de la fuente de iluminación que incide en el objeto y se dirigiría al punto (i, j) de la imagen generada o toda la imagen generada y puede tomar valores en el rango $0 < l(i, j) < \infty$. Finalmente, $r(i, j)$ es la fracción de la iluminación reflejada (no absorbida) por el objeto y puede tomar valores en el rango $0 < r(i, j) < 1$.

Es posible discretizar (digitalizar) el valor de intensidad mediante

$$n_L = 2^k \tag{2.3}$$

donde n_L son los niveles de intensidad y k es el número de bits usados para medir la intensidad una vez cuantificada la imagen. En este trabajo se utiliza $k = 8$ bits lo que significa que se tienen 256 niveles de intensidad.

Además en este proyecto se trabaja con:

Intensidad local se define la intensidad local como el valor discreto de nivel de gris que toma un píxel p en las cordenadas espaciales (i, j). Se trabaja con imágenes digitales de geles de electroforesis con intensidades en la escala de grises, que van de 0 a 255 (256 valores).

Intensidad global se define la intensidad global como el valor promedio de nivel de gris que toma una imagen digital.

2.2.2 Contraste

El contraste es una característica perceptiva del sistema visual humano asociada a diferencias de intensidad en una determinada región espacial. De forma general el contraste hace distinguible a un objeto de otros objetos y su fondo [51]. Ahora en el área de imágenes digitales se define contraste C como la diferencia relativa en intensidad entre un punto (píxel p) de una imagen I y sus alrededores (vecindad V o el resto de la imagen) [30], donde para la determinación de contraste en imágenes usualmente se tienen dos componentes [42]:

1. La diferencia entre dos intensidades, donde diferencias altas producen mayor contraste.
2. Alguna medición describiendo la adaptación del ojo humano, por ejemplo la intensidad promedio del objeto (imagen) en cuestión.

A partir de estos conceptos se definen las diferentes expresiones de medición de contraste en imágenes digitales y en este trabajo se utilizan las siguientes:

Contraste absoluto (C_A) es la diferencia normalizada entre la intensidad máxima L_{MAX} y miníma L_{MIN} de la región en estudio (toda la imágen o una parte de ésta) y se expresa matemáticamente por

$$C_A = \frac{L_{MAX} - L_{MIN}}{255} \tag{2.4}$$

Contraste de Michelson (C_M) en [51] se define el contraste de Michelson como la diferencia de la intensidad máxima L_{MAX} y miníma L_{MIN} entre la suma de las mismas en la región de estudio (toda la imágen o una parte de ésta) y es útil cuando se tienen patrones periodícos. Se expresa matemáticamente por

$$C_M = \frac{L_{MAX} - L_{MIN}}{L_{MAX} + L_{MIN}} \tag{2.5}$$

Contraste de Weber (C_W) en [42] se define el contraste de Weber como la diferencia entre la intensidad de un punto $L(i,j)$ y la intensidad media μ_L de la región de estudio (toda la imágen o una parte de ésta) dividido entre μ_L y se expresa matemáticamente por

$$C_W = \frac{L(i,j) - \mu_L}{\mu_L} \tag{2.6}$$

Contraste de intensidad (C_L) en [42] se define el contraste de intensidad como la diferencia de la intensidad máxima L_{MAX} y miníma L_{MIN}, entre la intensidad media μ_L de la región de estudio (toda la imágen o una parte de ésta) y se expresa matemáticamente por

$$C_L = \frac{L_{MAX} - L_{MIN}}{\mu_L} \tag{2.7}$$

Contraste RMS (C_{RMS}) en [51] se define matemáticamente el contraste RMS como

$$C_{RMS} = \sqrt{\frac{1}{MN-1}\sum_{i=0}^{m-1}\sum_{j=0}^{n-1}(L(i,j) - \mu_L)^2} \tag{2.8}$$

donde el C_{RMS} utiliza los valores de intensidad de los píxeles en la región de estudio (toda la imágen o una parte de ésta) y su media aritmética para proporcionar un único valor de contraste.

Además en este proyecto se trabaja con:

Contraste local opera en pequeñas áreas la imagen digital y su valor depende del tipo de expresión de medición de contraste utilizado.

Contraste global opera en toda la imagen digital y su valor depende del tipo de expresión de medición de contraste utilizado.

2.2.3 Ruido

Ruido es la contaminación del valor de un píxel de una imagen que produce una divergencia entre valor medido y el valor real de dicho píxel [30]. El ruido afecta directamente la percepción de calidad de una imagen digital y dificulta el análisis automatizado porque a mayor ruido, menor grado de relación entre los valores de los píxeles de la imagen y la realidad, lo que obliga a utilizar estimadores para inferir la apariencia real de los objetos.

La principal fuente de ruido en imágenes digitales se da en el proceso de su adquisión. Ésto se debe, principalmente, a procesos cuánticos asociados a la temperatura de los elementos de estado sólido empleados en la conversión de potencia lumínica a magnitudes eléctricas (CCD o CMOS) que a su vez será convertido en último término a valores digitales.

Algunos de los tipos de ruido más comunes presentes en imágenes digitales son:

Ruido blanco normal llamado así por la distribución estadística que lo describe, por eso también se le denomina ruido blanco gaussiano. El ruido blanco normal tiene un efecto general en toda la imagen, es decir, la intensidad de cada píxel de la imagen se ve alterada por una magnitud distribuida normalmente.

Ruido impulso (sal y pimienta) el ruido impulsonial por lo general es de tipo bipolar, donde los valores de la imagen se modifican o hacia blanco o hacia negro (de aquí su nombre sal y pimienta) distribuidos aleatoriamente por toda la imagen digital. Éste también puede ser de tipo unipolar si se presenta solo uno de los valores. A diferencia del ruido blanco normal el ruido impulsional tiene un efecto sobre un subconjunto del total de píxeles de la imagen.

2.2.4 Fusión de imágenes

La fusión de imagenes es un proceso que combina información de diferentes imágenes de una misma escena para formar una imagen con mayor calidad [33]. Por lo general se da énfasis a la información más relevante de cada imagen, según los dominios a mejorar. Para este trabajo estos dominios son el contraste y el ruido en imágenes digitales de geles de electroforesis, donde se toman imágenes a diferentes niveles de intensidad para un mismo gel de electroforesis con el objetivo de formar una imagen digital del gel analizado con mejor contraste y el menor ruido posible.

Antes de realizar la fusión para generar la imagen fusionada I_f se debe elegir el conjunto de las imágenes a fusionar, conocidas como las imágenes de entrada I_{in} del sistema de fusión $\mathscr{F}[\cdot]$ donde éstas deben ser seleccionadas de forma tal que que presenten valores de intensidad múltiples para captar detalles propios de cada imagen y mejorar la visualización final en una imagen con más detalles y valores de intensidad únicos [52].

En [33] se describe que el proceso de fusión puede ser realizado en tres niveles diferentes según lo que se requiera:

Nivel de píxel en este nivel los métodos de fusión operan directamente sobre las intensidades de los píxeles de las imágenes de entrada. Es útil cuando las imágenes son relativamente homogéneas; además, el procesamiento es rápido.

Nivel de característica en este nivel primero se extraen las características de las imágenes de entrada y mediante algún criterio se seleccionan las deseadas. Por último se realiza el proceso de fusión basado en las características seleccionadas. Este método proporciona la oportunidad de seleccionar qué características se desean mejorar en la fusión.

Nivel de decisión este nivel involucra la selección y clasificación de objetos en las imágenes fuente, donde los métodos de fusión pueden operar sobre objetos específicos. Es útil cuando las imágenes son muy heterogéneas y se desean fusionar solo partes de las imágenes de entrada.

A partir de este criterio los métodos de fusión seleccionados en los cuales se basan los algoritmos de fusion de imágenes implementados en este trabajo son:

- Fusión simple (SF: Simple Fusion)
- Fusión de exposición (EF: Exposure fusion)

Fusión simple (SF)

El método de fusión simple se presenta en [61] y el mismo consiste en obtener una imagen fusionada a partir de diferentes imágenes de entrada con distintos niveles de intensidad. Este sistema de fusión está dado por

$$I_f = \mathscr{F}_S[I_{in1}, I_{in2}, \ldots, I_{ink}] \tag{2.9}$$

donde la expresión matemática del sistema $\mathscr{F}_S[\cdot]$ para calcular la intensidad que se almacena en cada píxel de la imagen fusionada es

$$L_f(i,j) = \frac{1}{k}\sum_{x=1}^{k} L_{inx}(i,j) \tag{2.10}$$

donde k es el número de imágenes a fusionar.

2.2.5 Fusión de exposición (EF)

El método de fusión de exposición se presenta en [47] y el mismo consiste en obtener una imagen fusionada a partir de diferentes imágenes de entrada con distintos niveles de intensidad, donde se asignan pesos W a cada píxel según su valor de contraste C, saturación S y nivel de exposición E. En la tabla 2.1 se explica cada uno de estos paramétros del píxel y la forma matemática de obtener estos pesos es

$$W_{inx}(i,j) = (C_{inx}(i,j))^{w_C} \times (S_{inx}(i,j))^{w_S} \times (E_{inx}(i,j))^{w_E} \tag{2.11}$$

con los exponentes w_C, w_S y w_E con valores en el intervalo de 0 a 1 según el grado de influencia que se le desea dar a cada paramétro en el proceso de fusión.

Tabla 2.1: Paramétros relevantes del píxel para el algoritmo EF

Paramétro	Descripción	Valor
C	Diferencia de intensidad del píxel con sus alrededores	Depende del método de medición de contraste utilizado
S	Desviación estándar entre los canales de color del píxel	Depende del método de medición de saturación utilizado
E	Relevancia que se le toma al nivel de intensidad del píxel	Dado por $\exp\left\{-\frac{(L(i,j)-L_{central})^2}{2\sigma^2}\right\}$

Este sistema de fusión está dado por

$$I_f = \mathscr{F}_E[I_{in1}, I_{in2}, \ldots, I_{ink}] \tag{2.12}$$

donde la expresión matemática del sistema $\mathscr{F}_E[\cdot]$ para calcular la intensidad que se almacena en cada píxel de la imagen fusionada es

$$L_f(i,j) = \frac{1}{\sum_{x=1}^{k} W_{inx}(i,j)} \sum_{x=1}^{k} L_{inx}(i,j) W_{inx}(i,j) \tag{2.13}$$

con k es el número de imágenes a fusionar.

2.3 Detección de carriles y rectificación de imágenes

En la captura de las imágenes de geles de electroforesis se presentan dos tipos de distorsiones:

1. El primer tipo se debe al proceso de generación de los geles en sí, ya que las condiciones de campo eléctrico y temperatura generalmente no se mantienen constantes: al darse variaciones en estos parámetros se presentan desviaciones en los carriles.
2. El segundo tipo se debe al proceso de captura, entre las cuales se encuentran el ruido, incorrecto ajuste de la ganancia de la cámara y el bajo contraste

La principal manifestación del segundo tipo de distorsión es la deformación en forma de barril [57] debida al arreglo de lentes en el objetivo de la cámara. Esta se manifiesta como una contracción en las esquinas de la imagen.

Dentro del segundo tipo también se incluyen las distorsiones que se introducen debido a que la cámara no tiene su eje óptico perfectamente alineado con la perpendicular al plano del gel (distorsiones perspectivas) y a rotaciones entre ambos elementos.

Ambos tipos de distorsiones aumentan la complejidad de detección de los carriles en una imagen de gel, que de otro modo se limitaría a colocar líneas rectas de forma periódica sobre la imagen.

García [28] en su trabajo corrige las distorsiones mencionadas mediante el uso de una rejilla de calibración, la cual consiste en un cuadrícula ideal con dimensiones conocidas. Este método requiere obtener una imagen de la rejilla cada vez que las condiciones de captura cambien (posición de la cámara, ángulo, objetivo utilizado, etc.) para poder ajustar el modelo. Por esta razón se debe buscar un nuevo enfoque que sea independiente de una imagen de calibración. No se encontró otra referencia que corrigiera las distorsiones ópticas y detectara los carriles simultáneamente.

Otros trabajos de detección de carriles, como el desarrollado por Bailey y Christie [11], se basan en que existe una separación entre los carriles, la cual no existe si los geles son de los del tipo *diente de tiburón*, en los que la separación es eliminada para reducir costos en la generación de los geles. Además en ocasiones es necesaria la intervención del usuario para añadir o eliminar carriles, o en el peor de los casos como en el método propuesto por Glasbey et al. [29] se ubican todos los carriles manualmente.

2.3.1 Modelos activos de forma

Un Modelo Activo de Forma (ASM) es una técnica propuesta por Cootes et al. [18] que se basa en una la ubicación de n hitos etiquetados de forma única que se distribuyen en posiciones específicas de una silueta a representar. Mediante la ubicación de estos hitos en un conjunto de imágenes para entrenamiento y la obtención de estadísticas de la posición de los distintos puntos se deriva un Modelo de Distribución de Puntos (Point Distribution Models, PDMs). El PDM contiene el valor medio de las formas, así como parámetros que controlan las principales formas de variación de las mismas, obtenidos a través de análisis de componentes principales. La principal característica de esta técnica es la capacidad que tiene el modelo de deformarse pero solo siguiendo las formas típicas de variación encontradas en el conjunto de entrenamiento del modelo.

Para el caso particular de detección de carriles, un ASM se crea utilizando como cuerpo de entrenamiento las formas teóricas de distorsión tipo barril, dadas por

$$r_d(r_u) = r_u + ar_u^3 + br_u^5 + ... \tag{2.14}$$

con a y b factores de escala, r_d el radio distorsionado y r_u el radio no distorsionado.

2.3.2 Autocorrelación espacial

La correlación es un operador matemático utilizado como una medida de comparación entre dos señales. Dadas dos funciones reales $f(x)$ y $g(x)$ la correlación se define a través de la convolución como:

$$\phi_{fg}(x) = f(x) * g(-x) = \int_{-\infty}^{\infty} f(\tau)g(\tau + x)\, d\tau$$

Si $f(x) = g(x)$ a la operación anterior se le denomina auto-correlación, y se utiliza con frecuencia en la detección de periodicidades [54].

La función de autocorrelación es par, tiene su máximo absoluto en el origen, y si la función autocorrelada es periódica, tendrá máximos locales con el mismo periodo.

2.4 Estimación del efecto sonrisa

En la estimación del efecto sonrisa se hace uso del cálculo de "bordicidad", dada por medio de los pasos involucrados en el detector de bordes de Canny [16], usualmente basado en el cálculo de la magnitud del gradiente [30], pero empleando el ángulo para realizar un paso de eliminación de no-máximos. Además se hace uso de filtros de rango, en particular el filtro de máximos [57].

2.4.1 Difusión Anisotrópica

En [62] se plantea la comparación entre el proceso de difusión en imágenes con el proceso físico de que sucede para mantener homogenidad en un fluido. En el ambiente físico un proceso de difusión es aquel donde partículas son desplazadas de un lugar a otro para equilibrar cambios en la sustancia. Ejemplos de estos cambios son diferencias en la temperatura o en la concentración dentro de la sustancia. La ley de Flick describe esta difusión y es representada por la ecuación

$$j = -\mathbf{D} \cdot \nabla f \tag{2.15}$$

donde j es el flujo de sustancia, el operador ∇f calcula el cambio detectado y $\mathbf{D}$ es una matriz simétrica positiva que describe la relación entre j y ∇f conocida como coeficiente o tensor de difusión.

Ecuación de continuidad

En el proceso de difusión las partículas son transportadas, no creadas o destruidas, esto es reflejado en la ecuación de la continuidad

$$\partial_t f = -\mathrm{div}(j) \tag{2.16}$$

utilizada en este caso para la conservación de la materia. Uniendo (2.15) y (2.16) se obtiene la ecuación de difusión

$$\partial_t f = \mathrm{div}(\mathbf{D} \cdot \nabla f) \tag{2.17}$$

El coeficiente de difusión $\mathbf{D}$ determina el tipo de difusión que se realizará, siendo una difusión homogénea aquella que se realiza de la misma forma sobre todo el espacio, mientras que una no homogénea depende de la posición en el espacio. Además se puede clasificar la difusión como isotrópica o anisotrópica; la primera es aquella donde el flujo resultante

tiene la misma dirección que el cambio, mientras que en la segunda la dirección del flujo es diferente a la del gradiente. Un ejemplo de difusión isotrópica es la difusión gaussiana.

En cuanto a la difusión anisotrópica, permite utilizar información de posición, vecindad o derivadas de éstas para difuminar cada punto de acuerdo con la información del elemento. Un ejemplo de ella es la difusión con realce en la coherencia de la imagen (CED).

Difusión para realce de la coherencia de la imagen (CED)

La CED (*coherence enhanced diffusion*) es propuesta en [63], donde es definida como una difusión anisotrópica ya que genera el coeficiente de difusión a partir de la información presente en la imagen, no la posición únicamente. El utilizar dicha información permite encontrar las tendencias presentes en una región específica de la imagen y utilizar esa información para controlar la difusión.

Para obtener el coeficiente de difusión para realizar CED se debe obtener primero el coeficiente estructural, el cual esta dado por

$$\mathbf{J}_\rho = K_\rho * \nabla f_\sigma \nabla f_\sigma^T \tag{2.18}$$

con

$$K_\rho = \frac{1}{(2\pi\rho^2)^{m/2}} \cdot \exp\left(-\frac{|x|^2}{2\rho^2}\right) \tag{2.19}$$

La matriz $\mathbf{J}_\rho$ es el coeficiente estructural, el cual tiene la característica de ser una matriz semidefinida y simétrica. K_ρ es un filtro gaussiano de varianza ρ^2. f_σ es la imagen suavizada por un filtro gaussiano de varianza σ^2 y ∇f_σ es el gradiente de dicha imagen.

Se utiliza la imagen filtrada para eliminar el ruido presente en ella. La varianza σ^2 debe estar dimensionada de acuerdo a la imagen, ya que un valor muy pequeño de ella no elimina suficiente ruido generando resultados incorrectos en la lectura del flujo de la imagen, mientras que un valor grande de σ^2 puede borrar los bordes de la imagen reduciendo fuerza al flujo leído de ella.

El producto de los gradientes se utiliza con la finalidad de eliminar sus signos dejando una imagen que provee información de orientación del gradiente. Para dispersar la información de orientación obtenida de dicho producto se le aplica el filtro gaussiano de varianza ρ^2.

Los valores propios de la matriz $\mathbf{J}_\rho$ representan la dirección y la variación del gradiente. Como se explicó anteriormente de manera ortogonal a la dirección del gradiente se encuentra el borde. El conjunto de valores propios ortogonales a los valores propios de $\mathbf{J}_\rho$ dan la orientación con la menor variación en la imagen, el flujo natural de la imagen o también llamado la dirección de coherencia.

El coeficiente de difusión $\mathbf{D}$ debe permitir una difusión que siga la dirección de coherencia. Para ello se construye $\mathbf{D}$ de manera que tenga los mismos vectores propios de $\mathbf{J}_\rho$ pero sus

valores propios estarán dados por

$$\lambda_i = \alpha \quad \text{para } i \in \{1, \ldots, m-1\} \tag{2.20}$$

$$\lambda_m = \begin{cases} \alpha & \text{si } \kappa = 0, \\ \alpha + (1-\alpha)\exp\left(\frac{-C}{\kappa}\right) & \text{caso contrario} \end{cases} \tag{2.21}$$

donde m es el número de dimensiones a las que se va a aplicar el proceso de difusión; en el caso de imágenes $m = 2$. κ es la magnitud de la orientación de coherencia. α es una constante reguladora cuya función es mantener siempre definido el coeficiente de difusión y dar valor mínimo a la difusión en la dirección de coherencia, aún cuando la magnitud de la misma sea cero. C tiene función de umbral, así, si $\kappa \gg C \Rightarrow \lambda_m \approx 1$ pero si $\kappa \ll C \Rightarrow \lambda_m \approx \alpha$.

CED optimizada para invarianza rotacional La CED se basa en seguir el flujo de las imágenes. Es por ello que se debe procurar que al leer los flujos de la imagen la información no se vea afectada por las discretizaciones utilizadas. Este efecto no deseable sucede cuando se utiliza CED con un operador gradiente discretizado como el expresado por el kernel de Roberts o por diferencias simples [30, 53] en conjunto con valores altos en la discretización, volviendo al sistema inestable.

Es por ello que en [64] demuestran cómo utilizando los avances que se han obtenido en máscaras para el cálculo de derivadas se puede obtener una invarianza en la rotación y una mejora en la estabilidad de la CED.

Las máscaras que se proponen en [64] son

$$h_x = \frac{1}{32}\begin{bmatrix} -3 & 0 & 3 \\ -10 & 0 & 10 \\ -3 & 0 & 3 \end{bmatrix} \qquad h_y = \frac{1}{32}\begin{bmatrix} 3 & 10 & 3 \\ 0 & 0 & 0 \\ -3 & -10 & -3 \end{bmatrix} \tag{2.22}$$

para utilizarlas con la convolución con el fin de obtener los gradientes necesarios para aplicar una CED en la imagen. También en [64] se aclara que no sólo estas máscaras pueden conseguir una invarianza en la rotación, sino que todas aquellas máscaras que tengan un buen comportamiento en los ejes en cualquier dirección pueden ser utilizadas para conseguir este efecto. Además se afirma que con máscaras más grandes se puede tener una mejor invarianza en la rotación. Un ejemplo de máscara que cumple con lo anterior es la máscara tipo Sobel.

2.4.2 Dispersión con confiabilidad

La dispersión con confiabilidad fue expuesta por Antonio Aguilar en [1] (desarrollada en el contexto del proyecto anterior [7]) donde fue utilizada para acoplar la información de distorsión de las filas de la rejilla entre sí. Aguilar propone realizar una dispersión donde se pondere la capacidad de entregar información al resto de los elementos mediante la

confianza que posea el elemento. De esta manera los elementos adoptarán la información proveniente de aquellos que tengan una confianza superior a la suya, e ignorarán información de elementos poco confiables.

La dispersión con confiabilidad queda definida en [1] como

$$\underline{\tilde{\mathbf{y}}}(n) = \beta_n \sum_{k=-K}^{K} \underline{\mathbf{M}}(k)\underline{\mathbf{y}}(n-k)\underline{\mathbf{r}}(n-k) \tag{2.23}$$

donde $\underline{\tilde{\mathbf{y}}}$ es el vector resultante con la información influenciada por los puntos contenidos en él. El vector $\underline{\mathbf{y}}$ tiene la información a dispersar. El vector $\underline{\mathbf{r}}$ contiene la confianza de cada uno de los elementos de $\underline{\mathbf{y}}$. $\underline{\mathbf{M}}$ es la máscara con la que se quiere dispersar la información. La constante β_n asegura que $\beta_n \sum_{k=-K}^{K} \underline{\mathbf{M}}(k)\underline{\mathbf{y}}(n-k)\underline{\mathbf{r}}(n-k) = 1$ y esta dada por

$$\beta_n = \frac{1}{\sum_{k=-K}^{K} \underline{\mathbf{M}}(k)\underline{\mathbf{r}}(n-k)} \tag{2.24}$$

Para utilizar del concepto expuesto por Aguilar en sistemas de dos dimensiones se realiza la generalización la cual queda expresada como

$$\tilde{\mathbf{F}} = \beta[\mathbf{M} * (\mathbf{FC})] \tag{2.25}$$

con

$$\beta = \frac{1}{\mathbf{M} * \mathbf{C}} \tag{2.26}$$

donde $\mathbf{F}$ es la matriz a la cual se le desea dispersar la información contenida en sus elementos y en $\tilde{\mathbf{F}}$ es almacenado el resultado. $\mathbf{M}$ es la máscara con la que se desea dispersar la información, $\mathbf{C}$ es la matriz de confiabilidad que contiene la información sobre qué tan confiable es cada punto de $\mathbf{F}$ y $*$ es el operador de convolución.

2.5 Detección de bandas

2.5.1 Modelos de perfil de bandas

Los estudios realizados en el ajuste de forma de las bandas en geles de electroforesis se basan comúnmente en la distribución gaussiana [3]

$$g(x;\sigma,\mu) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}} \tag{2.27}$$

y esto se justifica en la investigación realizada en [44], la cual tiene como resultado que las bandas en electroforesis capilar en el análisis de ADN se aproximan a una distribución gaussiana, de donde se generaliza al caso de los geles con carriles más anchos.

En [3], se analiza el ajuste de bandas en geles de electroforesis utilizando la función lorentziana

$$g(x;\Gamma,\mu) = \frac{(\frac{\Gamma}{2})}{(x-\mu)^2 + (\frac{\Gamma}{2})^2} \tag{2.28}$$

y se obtiene que este proporciona una mejor aproximación, partiendo de que las bandas analizadas son más anchas que como propone el modelo gaussiano, ya que este último decae rápidamente.

Sin embargo, según [25] la distribución lorentziana no es utilizada comúnmente para modelado de bandas porque la cantidad de datos que deben tomarse en cuenta es mucho mayor dada la lenta caída de esta función. Además, si se realiza un análisis del espectro de esta distribución, la frecuencia de muestreo necesaria para obtener la información sin distorsión es considerablemente mayor que para el caso de la gaussiana, específicamente para obtener el $99,9\%$ de la información las frecuencias máximas que deben considerarse son $\omega_{max} = 3,29/\sigma$ para el caso de la función gaussiana y $\omega_{max} = 8,67/\sigma$ para el caso del lorentziano, esto en el caso ideal de no tener traslape de bandas. Esto conlleva a una mayor frecuencia de muestreo, lo que implica mayor resolución necesaria de las imágenes, que llevan a mayor tamaño y por tanto mayores tiempos de procesamiento que en el caso de utilizar el modelo gaussiano.

Por lo anterior se considera en este trabajo el uso de la distribución gaussiana como modelo para las bandas, puesto que implica realizar menos cálculos y considerar menos datos para cada banda, considerando teóricamente la distribución simétrica de estas para el análisis.

2.5.2 Estrategias para la ubicación de bandas

Tres tendencias se utilizan para resolver el problema de la ubicación automática de las bandas presentes en los geles de electroforesis:

En [17] la problemática se resuelve según el análisis molecular a realizar. La estrategia del método se basa en realizar un análisis de carriles orientados verticalmente, recorriendo con una plantilla en forma de "sonrisa" (forma de banda a ubicar) todo el carril, posicionando la plantilla en cada una de las filas de la imagen del carril, y calculando para cada fila en la cual la plantilla es ubicada el valor medio de los píxeles que la plantilla abarca, creando con cada uno de los valores medios obtenidos después de recorrer todo el carril una proyección. Se toma como criterio que la ubicación de las bandas se encuentra en los valles o mínimos locales de la proyección obtenida. Para facilitar la ubicación de los mínimos, inicialmente se ecualiza el histograma de intensidades [30], posteriormente se realiza una segmentación morfológica y por último la proyección se filtra con un filtro pasa bajas para eliminar las restantes irregularidades. El algoritmo finaliza con la ubicación de los mínimos locales de la proyección resultante. Sin embargo, esta estrategia parte de la premisa de que todas las bandas presentes en el carril han sufrido distorsión y presentan forma de "sonrisa" lo cual no siempre es cierto. Además depende en gran parte de la plantilla utilizada y no es capaz de detectar efectivamente las bandas en sectores con aglomeración de bandas ya que los identifica como una única banda.

Bajla et al. [12] destaca la problemática existente con el diseño de una estrategia de detección de bandas totalmente automática y propone una técnica basada en dos etapas,

que consideran la información del carril en sus dos dimensiones y la interacción con el usuario. En la primera etapa la imagen del carril es regularizada en intensidad utilizando un filtro GDD (Geometry Driven Diffusion), seguidamente con los carriles verticalmente orientados se aplica a la imagen un detector lineal de los límites horizontales de las bandas, el cual es un acumulador de diferencias de intensidad entre las filas, definido como:

$$D_i = \sum_i |I_{i+1,j} - I_{i,j}|$$

donde I_{ij},es el nivel de intensidad del píxel ubicado en la fila i y columna j. Posteriormente se realiza una ubicación de los máximos locales en D_i. El rectángulo creado entre cada par de máximos locales es considerado una banda. De esta forma finaliza la primera etapa y el usuario puede añadir o eliminar bandas a las actuales encontradas. Finalmente en la segunda etapa se utiliza el gradiente de la imagen en el vecindario de los píxeles que se detectaron como los bordes de las bandas con el objetivo de mejorar la ubicación de los límites. Para esto se forma un rectángulo que involucra el límite o borde dado por el gradiente, parte de la región considerada como fondo y parte de la región considerada como banda, obteniendo el indicador del límite final de la banda como la diferencia absoluta entre la media de la región del fondo y la media de la región de la banda. Debido a la regularización de intensidad realizada en la primera etapa del algoritmo, esta estrategia no permite realizar la ubicación de bandas que se encuentran en regiones del carril con bajo contraste, ya que bandas con bajos niveles de intensidad serán consideradas como fondo a pesar de poseer el perfil gaussiano de intensidad de una banda.

Por último en [29] y [37] se proponen dos estrategias basadas en la deconvolución mediante la estimación de parámetros utilizando el método estadístico de máxima verosimilitud (*maximum likelihood*). La primera de ellas para un carril verticalmente orientado, crea un carril con el promedio de intensidad de cada fila del carril original y posteriormente busca ajustarlo a uno de los perfiles de un genotipo previamente almacenado en una base de datos. No obstante, no siempre se cuenta con una base de datos que incluya el perfil del genotipo a analizar, lo cual limita la utilidad del método. La segunda busca ajustar la representación en 1D $y(s)$ de la secuencia de ADN con una sumatoria de funciones deltas de Dirac:

$$x(s) = A_0 + \sum_{j=1}^{p} A_j \delta(s - \tau_j) + e$$

considerando que la señal $y(s)$ se puede obtener mediante:

$$y(s) = x(s) * w(s)$$

donde $x(s)$ es la respuesta al impulso, la estrategia se enfoca en encontrar las posiciones centrales de las bandas τ_i y su respectiva amplitud A_j [7]. Sin embargo, se asume que no existen bandas en los extremos del carril y que existe ruido blanco e normalmente distribuido con baja varianza en todo el carril, lo cual no es cierto en las imágenes de los geles.

2.5.3 Estimación del ancho de las bandas

Para la estimación del ancho de las bandas se parte del supuesto de que el perfil de intensidades del carril presenta una distribución normal o gaussiana en los sectores de la imagen donde se localiza una banda presentando una desviación estándar común (la misma para todas las bandas). Para la estimación de la desviación estándar se hace uso del espacio de escalas (scale-space) [41, 39, 40].

En la figura 2.1 se ejemplifica el proceso de generación de el espacio de escalas. En este caso, por ser el perfil de intensidades una función unidimensional se utiliza un espacio de escalas unidimensional.

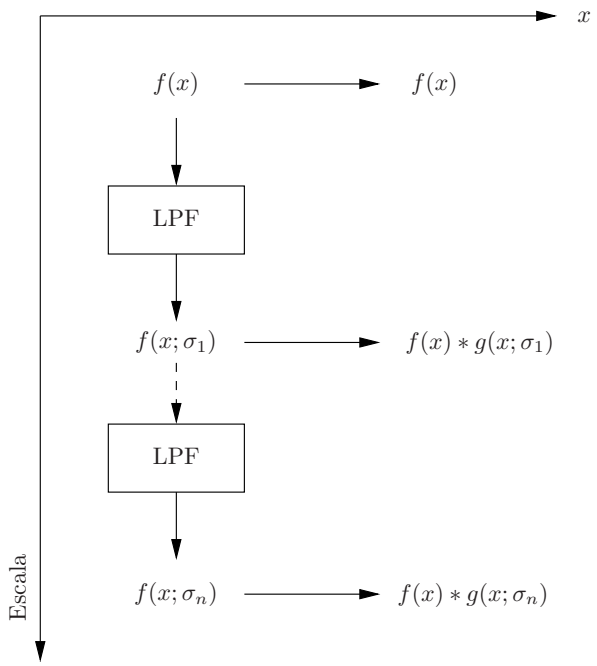


Figura 2.1: Representación de una función $f(x)$ en un espacio de escalas

El espacio de escalas se genera aplicando un filtro gaussiano a la imagen original mientras se varía de forma sucesiva incremental el parámetro de escala σ. Así, los niveles para una escala σ se calculan con

$$L(x;\sigma) = f(x) * g(x;\sigma) = \sum_{n=-\infty}^{\infty} f(n)g(x-n;\sigma) \tag{2.29}$$

Como punto de partida se tiene el perfil de intensidades de la imagen original $f(x)$, el cual corresponde a la escala $\sigma = 0$, o lo que es igual, a filtrar el perfil original con una función gaussiana de varianza cero $g(x;0)$ que a su vez equivale a un impulso de Dirac $\delta(x)$. Posteriormente se incrementa la escala σ y nuevamente se filtra el perfil original con la función gaussiana $g(x;\sigma_1)$, cuyo resultado corresponde a la siguiente fila de la imagen del espacio de escalas. Una vez hecho esto se incrementa nuevamente la escala y el resultado corresponde a la siguiente fila y este proceso se repite hasta tener una nueva imagen bidimensional formada a parir del perfil de intensidades original.

2.6 Evaluación multiobjetivo con frentes de Pareto

La evaluación multiobjetivo es una forma de evaluar algoritmos de procesamiento de imágenes. Fue introducido por Everingham *et al.* en [23] con la intención de estandarizar la forma de medir la eficiencia de algoritmos creados para la segmentación de imágenes.

Con la finalidad de evaluar se propone en [23] utilizar el conjunto de funciones de aptitud:

$$L(a_p, I) = \Psi(f_1(a_p, I), f_2(a_p, I), \ldots, f_N(a_p, I)) \tag{2.30}$$

donde a_p representa un algoritmo con p parámetros, I es el conjunto de imágenes definidas como deseables (un tipo de forma de evaluar los resultados), f_N son funciones que evaluan de manera individual alguna aptitud definida que debe cumplir el algoritmo. La función Ψ permite darle un peso a cada una de las medidas de aptitud sobre el comportamiento general del algoritmo.

Las funciones f_N deben ser definidas de manera que el crecimiento de ellas sea un comportamiento deseable. Además la función Ψ debe crecer cuando se produzca un aumento en los valores de todas las funciones de aptitud. El resultado de las funciones puede ser graficado en un espacio de N dimensiones donde cada una de las funciones representa una dimensión. Al evaluar el algoritmo utilizando el conjunto de parámetros p se obtendrá como resultado un valor para cada una de las funciones de aptitud, generando un punto en dicho espacio.

Un frente de Pareto lo que busca es eliminar aquellos puntos del espacio de N dimensiones de aptitud que no sean relevantes, clasificando como indeseable o no relevante aquel punto que ya exista en el espacio otro punto que tenga un mayor valor en la aptitud a-esima y menor valor en alguna las otras $N-1$ aptitudes. El conjunto de elementos que pertenecen al frente de Pareto esta dado por

$$\mathbb{P} = \{\langle a_p \in P_a, \underline{\mathbf{f}}(a_p, I)\rangle \mid \neg\exists a_q \in P_a : \underline{\mathbf{f}}(a_q, I) \succ \underline{\mathbf{f}}(a_p, I)\} \tag{2.31}$$

donde P_a es el espacio de aptitudes sobre el que se esta probando el algoritmo a, $\underline{\mathbf{f}}$ es el vector de pruebas $\underline{\mathbf{f}} = [f_1\ f_2\ \ldots\ f_n]^T$ y la relación $\succ$ es conocido como dominancia de Pareto y se define como

$$\underline{\mathbf{f}}(a_q, I) \succ \underline{\mathbf{f}}(a_p, I) \Leftrightarrow \forall i : f_i(a_q, I) > f_i(a_p, I) \wedge \exists i : f_i(a_q, I) > f_i(a_p, I) \tag{2.32}$$

Un ejemplo de un frente de Pareto es presentado en la Figura 2.2, donde los puntos rojos representan elementos dominados y la línea azul representa el frente de Pareto.

2.7 Arquitecturas para BD distribuidas

Diferentes autores han clasificado las distintas arquitecturas de bases de datos distribuidas de acuerdo con sus características de integración, control e independencia. Se seguirá la clasificación dada por Rahimi y Haug [56].

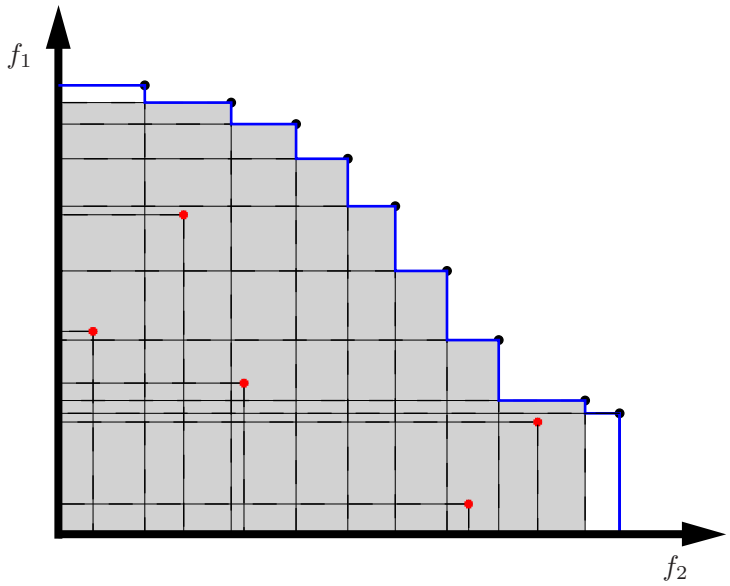


Figura 2.2: Ejemplo de frente de Pareto

2.7.1 Arquitectura tradicional (Distributed Data Base Management System, DDBMS)

En la arquitectura tradicional de bases de datos distribuidas se tienen varios Sistemas Administradores de Bases de Datos (SABD) conectados en una red. Dichos SABDs tienen como único propósito servir a la arquitectura distribuida como un todo. Esto es, no ofrecen funcionalidad independiente del sistema distribuido. Por medio de una o más interfaces los usuarios o aplicaciones pueden acceder a los datos contenidos en todas las bases de datos. Su arquitectura suele ser muy abierta porque usualmente requieren ser implementados en su totalidad con el fin de proveer la funcionalidad integrada que les caracteriza. En resumen, el sistema distribuido tiene control total de los esquemas de datos y de los datos.

Entre sus ventajas se cuenta que proveen un modelo unificado para todos los datos que se manejan en el sistema. Además permite una implementación afinada a las necesidades particulares de la organización. Finalmente, dado el nivel de unificación de los datos y modelos, no hay ni datos ni estructuras ocultas al sistema distribuido total.

Como desventaja de esta arquitectura se puede señalar que su implementación es muy laboriosa por lo que se debe reducir su funcionalidad para hacerla factible. Otra desventaja muy importante es que requiere un control muy rígido que le quita independencia a los nodos participantes, lo cual puede ir en contra de las necesidades organizacionales de esos participantes.

Debido al control total requerido por esta arquitectura, se considera que es poco apropiada para la aplicación de imágenes de geles propuesta. Se prevé que el sistema distribuido estaría formado por nodos participantes que están dispuestos a compartir algunos de sus datos. Sin embargo, las necesidades de procesamiento y control de esos participantes

no se pueden subordinar a un planteamiento externo global. Cada participante puede requerir de información adicional de interés muy particular. Además los nodos deben poder establecer limitaciones sobre la información que van a compartir.

2.7.2 Arquitectura federada

En la arquitectura federada de bases de datos, Federated Data Base Management Systems (FDBMS), se tienen varios SABDs conectados en red. En este caso, sin embargo, cada SABD es un sistema autónomo que ofrece servicios locales además de los servicios que forman parte de la arquitectura distribuida total.

Los diferentes nodos participantes pueden usar diferentes modelos de datos y organizaciones físicas. Debido a esta independencia, no es necesario implementar el sistema distribuido desde cero ya que los sistemas locales pueden ser sistemas generales de bases de datos ya disponibles. Los módulos que implementan los servicios distribuidos sí deben implementarse desde cero puesto que deben ligar la funcionalidad de sistemas que podrían ser muy distintos.

La independencia de los nodos permite que haya datos y esquemas de datos que son estrictamente para uso local y por lo tanto invisibles para el sistema distribuido. Esa misma independencia reduce el control del sistema distribuido por lo que lo usual es que los datos distribuidos sean de solo lectura. La capacidad para actualizar los datos y para velar por su integridad es usualmente reservada para aplicaciones locales externas al sistema distribuido.

Una ventaja de esta arquitectura es que permite una gran flexibilidad en la distribución de datos y en el uso de sistemas administradores de bases de datos. Además, no requieren implementar funcionalidad de bajo nivel porque dicha funcionalidad es provista por los sistemas locales. Finalmente, otra ventaja es que permite que los nodos locales dispongan de datos y de esquemas de datos propios para sus necesidades particulares.

Por otro lado, esta arquitectura tiene la desventaja de que puede ser difícil implementar complejos mecanismos de control distribuido como transacciones. Va a depender de cuán abiertos sean los sistemas locales para poder acceder a mecanismos de bajo nivel. Otra desventaja es que debido a lo reducido del control distribuido, los cambios en datos y esquemas no pueden ser hechos por una autoridad centralizada puesto que pueden afectar las aplicaciones particulares de cada nodo participante.

Esta arquitectura se ajusta bien a las necesidades de las organizaciones que quieren compartir datos de geles. En este caso, se desea alcanzar un nivel de estandarización que permita intercambiar datos, pero a la vez mantener la independencia que permita atender las necesidades particulares de cada organización. De hecho ya se han definido estándares para incorporarse a redes de bases de datos federadas [10].

2.7.3 Arquitectura cooperativa

En esta arquitectura, la motivación principal de los nodos participantes es compartir contenido; usualmente se comparte música u otro tipo de contenido multimedial. Mientras unos nodos participantes proveen contenido, otros nodos consumen contenido y hay terceros que sirven de intermediarios. Otra característica es que los participantes varían constantemente por lo que están solo ligeramente afiliados al sistema distribuido total. Se trata de ambientes muy dinámicos.

Esta arquitectura es muy resistente y eficiente a la hora de compartir contenido. No hay garantías, sin embargo de que el contenido esté disponible en un momento dado.

Esta arquitectura no es adecuada para organizaciones que quieren compartir datos de geles. La razón es que al estar los proveedores solo ligeramente ligados al sistema distribuido hay un menor compromiso en proveer su contenido. Para el caso multimedial, esto en general no es problema ya que el mismo contenido puede ser provisto por otros nodos participantes, pero en el caso de geles se espera que la gran mayoría de las organizaciones tengan un alto porcentaje de contenido que solo sea provisto por cada una de ellas.

2.7.4 Arquitectura entre pares (peer-to-peer)

Es una arquitectura muy similar a la arquitectura cooperativa. Su diferencia radica en que enfatiza el hecho de que todos los participantes son pares (peers) iguales, de modo que cumplen simultáneamente con los roles de consumidor, proveedor.

Debido a su similitud con la arquitectura anterior comparte sus ventajas y desventajas. Por un lado es muy resistente y eficiente, pero por otro lado al estar los nodos solo ligeramente ligados al sistema no proveen garantía alguna de participación. Para organizaciones interesadas en compartir información sobre geles se requiere un compromiso de participación más firme.

2.8 Consideraciones generales sobre minería de datos

La minería de datos es definida como el proceso de descubrir patrones en datos; el proceso puede ser automático, o más frecuentemente, semi-automático, y los patrones descubiertos deben ser significativos y representar una ventaja para quién los descubra [65].

La minería de datos en muchos casos surge como una continuación natural a los procesos típicos de construcción de un data warehouse. Esta construcción típicamente incluye etapas en las que los datos son depurados de errores, para ser luego integrados si provienen de múltiples fuentes. Tras lo cual son seleccionados y transformados para convertirlos a un formato susceptible de análisis. La labor de minería de datos propiamente dicha consiste en aplicar procesos inteligentes de extracción de patrones a la información almacenada

por las etapas anteriores. Luego se deben evaluar los patrones extraídos, para finalmente producir algún tipo de visualización del conocimiento.

La extracción de patrones en la minería de datos se realiza por medio de varias técnicas de análisis:

Descripción de clase: proveer una sumarización de un conjunto de datos que permita distinguirlos de otros conjuntos de datos.

Asociaciones: encontrar dentro de un conjunto de datos asociaciones o relaciones de la forma $X \rightarrow Y$ que se interpreta como un indicación de que los casos en que X se cumple muy probablemente también cumplan Y.

Clasificación: Se desea poder predecir las clases a las que pertenecen unos objetos; se parte de conjunto de entrenamiento, el cual consiste de objetos cuyas clases se conocen de antemano, y a partir de ese conjunto se construye un modelo para determinar las clases de los objetos de acuerdo con las características de cada objeto. Los modelos generados suelen tomar la forma de árboles de conjunto o de reglas de clasificación.

Predictores: Encontrar un conjunto de atributos en un conjunto de objetos que permita predecir los posibles valores de otro atributo.

Agrupamiento (clustering): Dado un criterio de similitud, identificar conjuntos de objetos que sean similares unos con otros de acuerdo con dicho criterio.

Análisis de series de tiempo: consiste en tomar un conjunto de datos serializados en el tiempo y analizarlos para encontrar regularidades interesantes como son secuencias similares o periodicidades.

2.8.1 Facilidades de minería de datos en otras herramientas de manejo de imágenes de geles

El análisis cluster y la clasificación son operaciones de minería de datos muy importantes para el análisis de imágenes de geles. Es fundamental para las herramientas proveer dicha funcionalidad. Por ejemplo, `Phoretix 1D pro` [38] es una herramienta para el análisis de imágenes de geles que incluye un módulo que permite hacer clustering de carriles y producir dendogramas que permiten estudiar las relaciones entre carriles seleccionados de la base de datos; además permite identificar y clasificar muestras desconocidas con respecto a una biblioteca de muestras de referencia previamente identificadas. El análisis cluster también es provisto por `Mascot Integra` [46] la cual es una herramienta que provee los protocolos experimentales usados en investigación proteómica, incluyendo geles 1D.

2.8.2 Software libre disponible para hacer Data Mining

Existe una gran cantidad de herramientas de software disponibles para hacer minería de datos. De hecho, se han elaborado muchas listas de esas herramientas. Por ejemplo, la lista que presenta `KDnuggets` [36] contiene 67 herramientas comerciales y 18 herramientas libres. El proyecto requiere utilizar software libre, de modo que se buscó una herramienta de este tipo que fuera poderosa, flexible y madura. La herramienta `Weka` (Waikato Environment for Knowledge Analysis) [45] cumple a plenitud con todas esas características. Es una herramienta muy destacada en las listas, y además incluye la funcionalidad de clustering y clasificación que interesa ofrecer en el módulo de Data Mining.

Weka implementa en Java un conjunto de algoritmos para realizar tareas típicas de minería de datos. Provee mecanismos para leer directamente los datos, o hacerlo por medio de código Java provisto por el usuario. Weka permite pre-procesar los datos, así como aplicar clasificación, regresión, análisis cluster, generación de reglas de asociación. También provee herramientas de visualización de los resultados.

Weka posee un API que permite acceder a su funcionalidad de minería de datos desde una aplicación de usuario. Como Weka está programado en Java es fácilmente integrable a la arquitectura de aplicación web de este proyecto. Solo se requiere agregar un servidor `Apache Tomcat` [60] para proveer el acceso a Java servlets que implementan el módulo de minería de datos.

Capítulo 3

Materiales y Métodos

Se presentan en este capítulo varias secciones relacionadas con el Sistema de Análisis de Imágenes (figura 1.2): la captura de la imagen (estudiada particularmente en la tesis de Bryant Álvarez [9]), mejoras en los algoritmos de detección de carriles (estudiado con la tesis de Pablo Barrantes [13] y en la corrección del efecto sonrisa (profundizado en la tesis de Pedro Alpízar) [5]. Se presentan además los avances en los algoritmos automáticos de detección de bandas en un carril (tesis de David Soto [59], Edison Fernández [26] y de Randall Esquivel [22]).

Con respecto al Sistema de Base de Datos (figura 1.2) se presentan los diseños realizados a partir de la sección 3.7. Finalmente, se presentan los métodos utilizados para la generación de geles DGGE en la sección 3.9.

3.1 Mejora de calidad de imagen desde la captura

En el procesamiento de imágenes se aprovecha optimizar la captura de las imágenes para reducir ruido y aumentar la cantidad de información de interés registrada en las imagenes en estudio. Con la tesis de Bryant Álvarez [9] se exploró la generación de imagénes de alto rango dinámico a partir de la captura de varias imagenes con diferentes parametrizaciones de la cámara.

3.1.1 Factores controlables en la captura

La calidad de una imagen depende de las características electrónicas de la etapa de adquisición, donde es deber del diseñador del sistema establecer las condiciones necesarias del proceso de formación de imágenes, lo que incluye configuración de la escena, características de la iluminación, etc., así como de conocer y definir los parámetros necesarios de la cámara digital para obtener la mejor calidad posible de una imagen digital. En el sistema diseñado se controlan tres diferentes parámetros de la cámara:

Brillo (*brightness*) este parámetro altera el nivel de intensidad total de toda la imagen, variando el nivel de intensidad individual de cada pixel, por lo que el brillo simplemente representa un "offset" en los niveles de gris de la imagen.

Ganancia (*gain*) este parámetro determina el nivel de amplificación de la señal de salida obtenida por el sensor de la cámara. Un incremento en la amplificación incrementa el contraste, sin embargo, produce imágenes más ruidosas.

Obturación (*shutter*) este parámetro determina el tiempo de exposición del sensor de la cámara, donde un nivel bajo de obturación produce un tiempo de exposición corto, lo que genera imágenes oscuras, mientras un nivel de obturación alto captura imágenes claras debido a que le llega luz durante más tiempo al elemento sensible

Se emplean medidas cuantitavas para medir la calidad de la imagen a través del contraste y el nivel de ruido estimado (NRE). Se presentan a continuación las mediciones de calidad empleadas para las imágenes de geles de electroforesis capturadas con diferentes parámetros. Con estadísticas de las mediciones de calidad en diferentes imágenes de geles se plantea la optimización del sistema de adquisición. Con el optimizador realizado, es posible seleccionar imágenes con baja intensidad luminosa (oscuras), media intensidad luminosa (medias), alta intensidad luminosa (claras), lo que facilita la selección de imágenes para la fusión, debido a que cada una de las imágenes con diferentes niveles de intensidad presenta características óptimas de contraste y ruido en diferentes zonas de la imagen. Al fusionarse las imágenes dan origen a otra con mayor índice de calidad.

3.1.2 Algoritmos de medición

Los algoritmos de medición utilizados se dividen en tres grupos:

- Algoritmos de medición de intensidad
- Algoritmos de medición de contraste
- Algoritmo de estimación de ruido

Algoritmos de medición de intensidad

Estos algoritmos se utilizan para medir el nivel de intensidad luminosa de las imágenes digitales y así poder clasificarlas en imágenes "claras", "medias" o "oscuras" según su valor. Los algoritmos de intensidad se separan en locales y globales según la información que proporcionan.

Intensidad local La intensidad local $L(i, j)$ en una imagen digital de gel de electroforesis es el valor discreto de nivel de gris que toma un píxel p en las coordenadas espaciales (i, j), donde la escala de grises tiene 256 valores discretos que van de 0 a 255. Por lo tanto, en este trabajo se obtiene la intensidad local de un píxel en la imagen digital I leyendo el

valor almacenado en las coordenadas espaciales (i, j) de la matrix de la imagen, obtenida mediante el sensor electrónico de la cámara.

Intensidad global La intensidad global en una imagen digital de gel de electroforesis es el valor promedio de nivel de gris, donde la escala de grises va de 0 a 255. Por lo tanto, en este proyecto se obtiene la intensidad global de una imagen digital mediante las siguientes expresiones:

Media aritmética de intensidad (μ_L) es el valor obtenido al sumar todos los píxeles p de la imagen y dividir el resultado entre el tamaño de la imagen digital $M \times N$, su expresión matemática es

$$\mu_L = \frac{1}{M \times N} \sum_{i=0}^{m-1} \sum_{j=0}^{n-1} L(i, j) \tag{3.1}$$

Mediana de intensidad (Me_L) es el valor que ocupa el lugar central de todos los pixeles en la imagen cuando éstos están ordenados de menor a mayor.

Algoritmos de medición de contraste

Estos algoritmos de medición se utilizan como un factor de calidad en las imágenes de geles de electroforesis, debido a que se busca cuantificar alguna característica que corresponda a la percepción del visual humana en cuanto a la diferencia relativa en intensidad entre un punto (píxel p) de una imagen I y sus alrededores (vecindad V o el resto de la imagen) [30]. De aquí que no exista una sola forma de cuantificar el contraste y por lo tanto, en este trabajo se utilizan diversos algoritmos para que evaluar las ventajas y desventajas de cada método. Por otra parte, los algoritmos se separan en locales y globales según actúen sobre áreas de la imagen o sobre la imagen completa respectivamente.

Contraste local El contraste local $\mathbf{C_{lo}}$ en una imagen se define como el grado de contraste en los límites de los detalles que tienen diferencias lumínicas. Cuanto más contraste haya en los límites entre una zona oscura y otra más clara, mayor será el contraste local y mayor será la percepción de la calidad de la imagen [35]. Con los algoritmos de contraste local se mide el contraste en pequeñas áreas de la imagen, lo que permite detectar cuáles zonas específicas de la imagen son en las que se necesitan mejorar el nivel de calidad para que el sistema de percepción visual humana y el sistema computacional sean capaces de distinguir más detalles.

Los algoritmos de contraste local generan matrices de contraste $\mathbf{C}$ de igual tamaño que la imagen digital, donde cada píxel p de la imagen I tiene su valor de contraste local asociado en la posición (i, j). Para esto se aplican operaciones de vecindad sobre pequeñas áreas de la imagen original con el fin de obtener las diferentes matrices de contraste. El tamaño

de estas áreas está dado por el tamaño de la vecindad $V_8(p)$ (rejilla cuadrada con pixel central p y vecinos q) seleccionado, donde los valores de contraste se van almacenando en la matrix de contraste, exactamente en la posición del píxel p y no en la posición de sus píxeles vecinos q. Esta operación se realiza desde el píxel $(0, 0)$ (primer píxel de la imagen) hasta el píxel $(M - 1, N - 1)$ (último píxel de la imagen). Se ofrecen cinco opciones de contraste: absoluto, de Michelson, de Weber, de intensidad y RMS en sus versiones locales (ver sección 2.2.2).

Contraste global El contraste global C_G en una imagen digital de gel de electroforesis corresponde a la diferencia total de valores extremos en ésta y proporciona una medición general de la calidad de la misma. Se puede seleccionar entre las 5 variables globales presentadas en la sección 2.2.2.

Algoritmo de estimación de ruido

El ruido producido por la cámara CCD no es simplemente aditivo, sino dependiente del nivel de intensidad de la imagen [27]. Por esta razón se diseña el algoritmo de estimación que considera la imagen como una señal con áreas homogéneas de intensidad contaminadas con ruido. En la figura 3.1 se muestra el algoritmo de estimación diseñado e implementado para el proyecto en [9], que consta fundamentalmente de las etapas:

1. Detección de bordes
2. Erosión
3. Áreas homogéneas
4. Estimación del nivel de ruido

Detección de bordes Los bordes de una imagen digital se definen como transiciones entre dos regiones de niveles de gris significativamente distintos. Suministran información sobre las fronteras de los objetos y se utilizan para segmentar la imagen, reconocer objetos, etc. [30]. Aquí se calculan y detectan los bordes de la imagen de entrada, para separar las áreas homogéneas.

Por otra parte en el área de procesamiento de imágenes digitales, la detección de bordes es una operación conocida. Para efectos de este trabajo se seleccionó el algoritmo de Canny [16] que es un algoritmo estándar basado en la optimización de un criterio de "calidad de borde" definido en términos de gradientes, supresión de no-máximos y umbralización.

Erosión En la etapa de detección de bordes se aplica el algoritmo Canny y se genera una imagen de tipo binario con los bordes representando el 0 (color negro) y el fondo el 1 (color blanco). En esta etapa se realiza la operación morfológica de erosión con el objetivo de expandir los bordes y obtener áreas de la imagen de entrada más homogéneas para la siguiente etapa del algoritmo de medición de ruido. La erosión [30] es un método

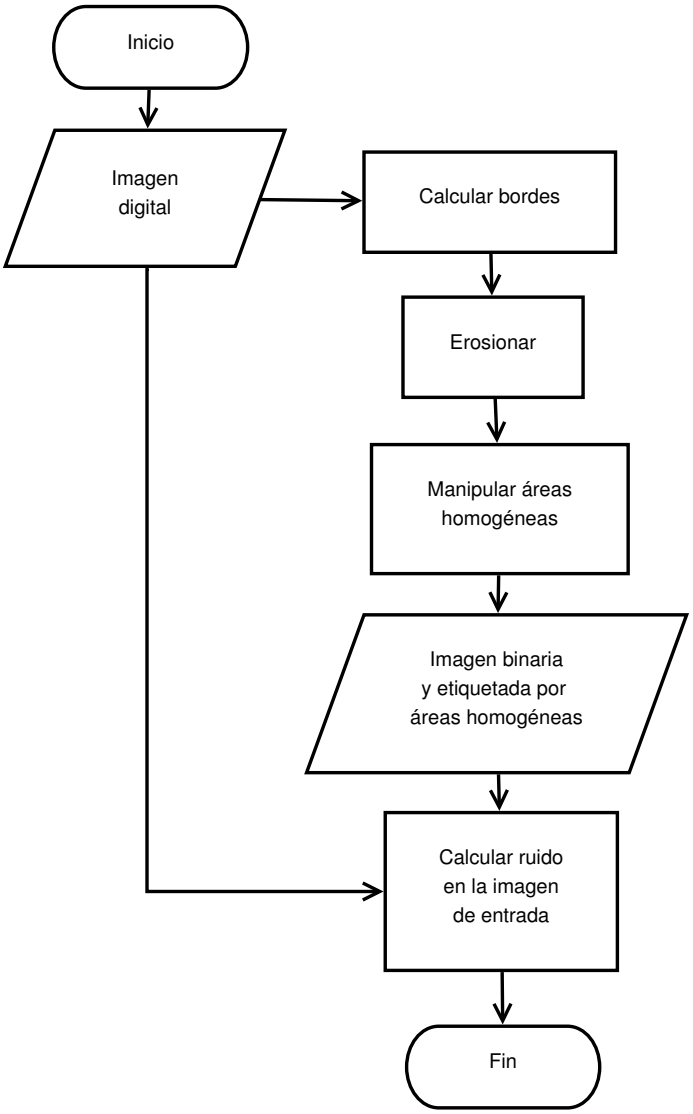


Figura 3.1: Diagrama de flujo del cálculo de ruido en imágenes digitales

del área de procesamiento de imágenes que aplica un elemento estructural (dado por el tamaño y la forma de la vecindad de un pixel seleccionada) a la imagen de entrada, sin cambiar el tamaño de la imagen de salida.

Áreas homogéneas En esta etapa se determinan, clasifican y etiquetan las áreas homogéneas en la imagen procesada proveniente de la etapa de erosión. Para realizar está operación se utilizó un algoritmo de etiquetación rápida [34], el cual detecta componentes conectados entre píxeles en imagenes binarias, donde los bordes son los separadores entre áreas homogéneas cuando encierran segmentos completos.

Estimación del nivel de ruido Esta etapa recibe la imagen binaria y etiquetada obtenida mediante todas las etapas anteriores. Se parte del supuesto que esta imagen aproxima los segmentos con valores de intensidad homogénea en la imagen original, por

lo que diferencias en los valores de píxel en estas áreas homogéneas se deben al efecto del ruido producido principalmente por la cámara CCD.

Para estimar el ruido en la imagen original se realiza el cálculo de la diferencia de intensidades de cada píxel de la imagen original con la respectiva media aritmética del segmento de la imagen a la que pertenece este píxel, excluyendo los píxeles que determinan bordes expandidos en la imagen original. Por otra parte, puesto que esta diferencia de intensidades puede ser negativa, se utiliza

$$\text{NRE} = \sqrt{\frac{1}{\sum_{SH=1}^{k} T_{SH}} \sum_{SH=1}^{k} \sum_{p=1}^{T_{SH}} (L_{SH}(p) - \mu_{SH})^2} \tag{3.2}$$

expresión matématica similar a la desviación extándar, con la diferencia de que existe una única media aritmética μ_{SH} para cada segmento homogéneo, obtenida mediante

$$\mu_{SH} = \frac{1}{T_{SH}} \sum_{p=1}^{T_{SH}} L_{SH}(p) \tag{3.3}$$

donde T_{SH} es el tamaño (número de píxeles) del segmento homogéneo utilizado y k el número de segmentos homogéneos. El NRE obtenido corresponde a la desviación que produce el ruido en los píxeles de la imagen de entrada.

3.1.3 Optimización del sistema de adquisición

El sistema de adquisición de imágenes digitales propuesto está formado por la cámara digital y un optimizador donde se determinan los parámetros de brillo B, ganancia G y obturación O de la cámara para obtener imágenes con diferentes niveles de intensidad específicamente en tres rangos: baja intensidad luminosa (imagen oscura I_o), media intensidad luminosa (imagen media I_m) y alta intensidad luminosa (imagen clara I_c); con el mayor contraste y menor ruido posible.

El sistema de adquisición es un sistema electrónico que depende tanto de las entradas como de las salidas (lazo cerrado) con las siguientes características:

- MIMO (Múltiples entradas, múltiples salidas)
- No lineal

El sistema presenta perturbaciones como la iluminación, temperatura, tipo de sensor de la cámara, modo de transmisión de datos, entre otras [43]. Por eso se utilizan técnicas de manipulación aproximadas que presentan rangos de equilibrio.

En la figura 3.2 se muestra el sistema de optimización propuesto para la adquisición de imágenes digitales.

El manejo de la cámara debe ser capaz de auto-ajustar los parámetros manipulables (B, G y O) y además evaluar los resultados para saber si se obtuvo la imagen deseada. Para ésto se necesita contar con una base de datos en el optimizador, adquirida previamente

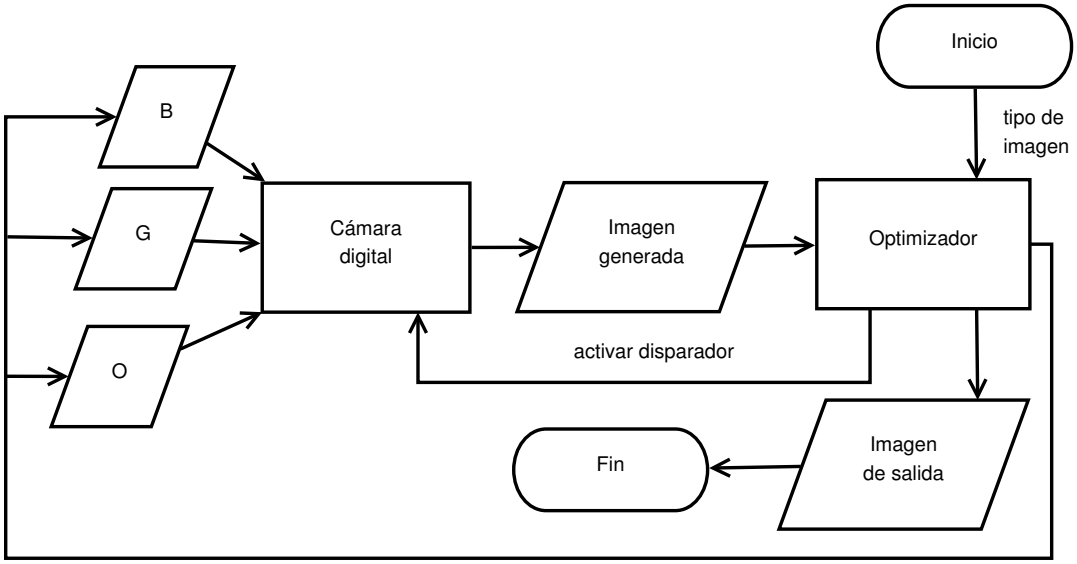


Figura 3.2: Diagrama del sistema de adquisición optimizado de imágenes digitales

mediante la simulación de situaciones específicas. Se divide el diseño del optimizador en dos etapas;

- Análisis estadístico del comportamiento del sistema
- Evaluación y manipulación

Análisis estadístico

En esta etapa se generan imágenes digitales mediante simulaciones para obtener las condiciones necesarias para la etapa de manipulación y evaluación. El procedimiento a seguir es el siguiente:

1. Se generan imágenes variando cada uno de los parámetros a utilizar de la cámara (B, G y O) individualmente y grupalmente, a diferentes niveles de perturbación, como por ejemplo, imágenes generadas a diferentes horas del día y a niveles de iluminación artificial diferentes.

2. Después, se mide el efecto de estos parámetros sobre las características a mejorar (contraste y NRE) en las imágenes generadas.

3. Posteriormente, se obtienen relaciones entre las entradas y salidas del sistema en base a los resultados obtenidos en el punto anterior.

4. Finalmente, debido a que en el optimizador lo que se desea es obtener imágenes claras, medias y oscuras con el mejor contraste y NRE posibles, se asigna un rango de intensidades globales para lo que es detectado por el sistema como imágenes claras, medias y oscuras, considerando que éstas no pueden ser ni extremadamente claras, ni extremadamente oscuras.

Etapa de evaluación y manipulación

En esta etapa se toman en cuenta las condiciones obtenidas en el análisis estadístico de las simulaciones realizadas y a partir de estas se diseña la evaluación de imágenes y manejo de la cámara digital.

Con el análisis estadístico del sistema de adquisición de este proyecto, se obtuvieron las siguientes condiciones para el diseño del optimizador:

1. El paramétro del brillo B en la cámara no realiza ningún efecto significativo sobre el ruido y el contraste en imágenes, por lo que se puede eliminar de las variables a manipular.

2. O y G se comportan similar, ambos aumentan la intensidad global de la imagen cuando se aumenta su valor.

3. El ruido incrementa con la ganancia, mientras que la tasa de captura disminuye con la obturación, por lo que para este sistema se utilizan G bajas (en el intervalo $[350, 500]$ para la cámara utilizada) y O es la variable manipulada, puesto que lo que se desea son imágenes y no video.

4. El nivel de iluminación varía el valor de los paramétros necesarios para obtener intensidades globales iguales en las imágenes.

5. El ruido y el contraste se comportan de manera similar, sin embargo, ganancias bajas disminuyen el ruido pero no así el contraste.

6. Se define el rango de imágenes en la escala de grises como sigue, I_c en el intervalo $[180, 220]$, I_m en el intervalo $[130, 170]$, I_o en el intervalo $[80, 120]$.

Tomando en cuenta estas condiciones se diseña el manejo automático del sistema de adquisición mediante el algoritmo de la figura 3.3. La imagen de inicialización depende del tipo de imagen (I_o, I_m o I_c) a obtener por el optimizador, donde los valores de B, O y G iniciales para capturarla se estimaron en el análisis estadístico para cada nivel de intensidad requerido. Después, con base en la intensidad global de la imagen de inicialización se selecciona entre las condiciones de poca iluminación o iluminación "normal" para generar diez imágenes con diferentes valores de B, O y G estimados y por último seleccinar de este conjunto la que posee mejor valor de contraste global y menor nivel de ruido estimado.

3.1.4 Algoritmos de fusión de imágenes

Estos algoritmos permiten mejorar el nivel de contraste y el NRE, mediante la fusión de imágenes a diferentes niveles de intensidad de una misma escena, con el objetivo de generar una imágen única de mejor calidad (figura 3.4).

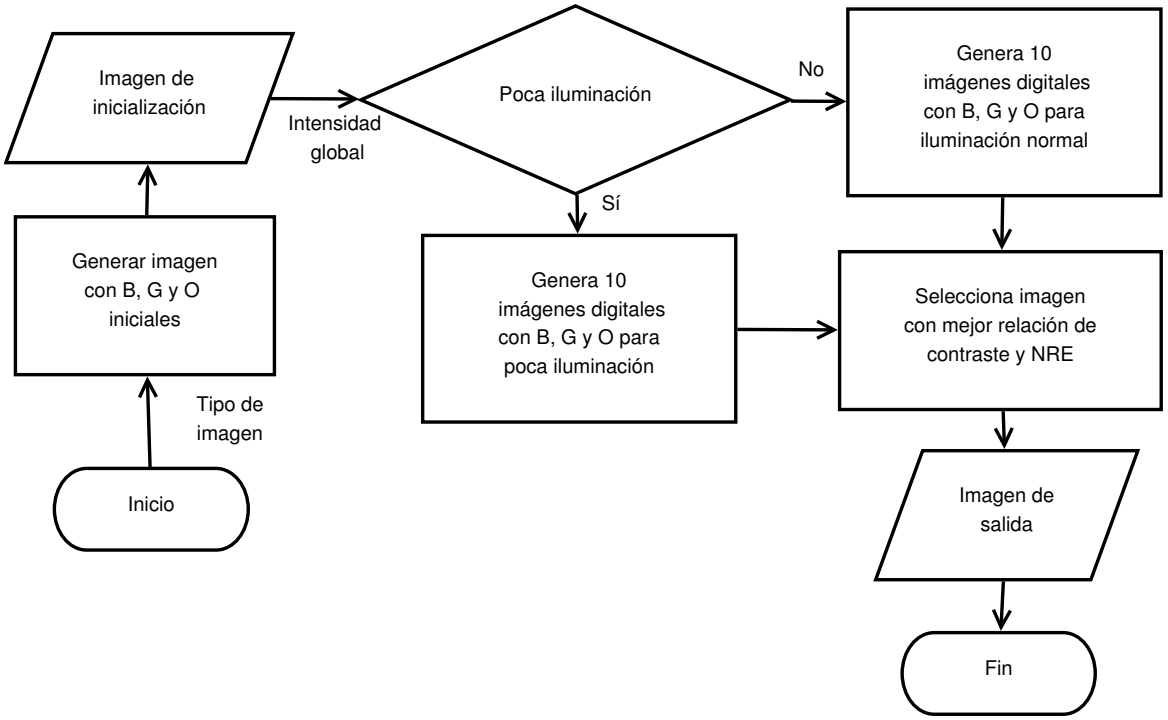


Figura 3.3: Diagrama de flujo del manejo para el sistema de adquisición

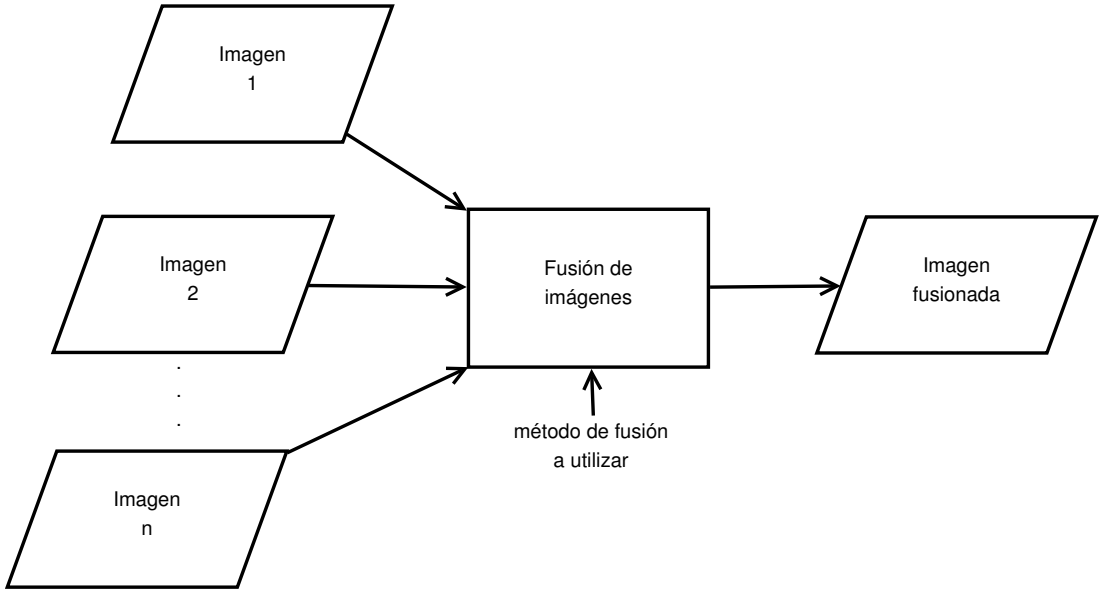


Figura 3.4: Diagrama de fusión de imágenes digitales

Para el diseño de los algoritmos de fusión implementados se toman en cuenta las siguientes consideraciones [33]:

1. El proceso de fusión debe preservar toda la información relevante de las imágenes de entrada en la imagen fusionada.
2. La técnica de fusión no debe provocar ninguna inconsistencia con la realidad en la imagen final.
3. Efectos indeseados (bajo contraste) y el ruido deben ser suprimidos al máximo.

Además como se mencionó en el marco teórico el proceso de fusión puede ser realizado en tres niveles diferentes y en este proyecto se selecciona un algoritmo a nivel de píxel y otro a nivel de característica para comparar resultados. No se seleccionó ningún algoritmo a nivel de decisión debido a que las imágenes digitales de geles de electroforesis no presentan objetos diferentes y se desean mejorar sus propiedades en toda la imagen. Los algoritmos seleccionados son:

- Algoritmos de fusión simple (nivel de píxel)
- Algoritmo de fusión de exposición (nivel de característica)

Algoritmo de fusión simple (SF)

Este algoritmo utiliza el método de fusión simple presentado en el marco teórico (sección 2.2.4), donde se obtiene una imagen fusionada a partir de diferentes imágenes de entrada con distintos niveles de intensidad (figura 3.5).

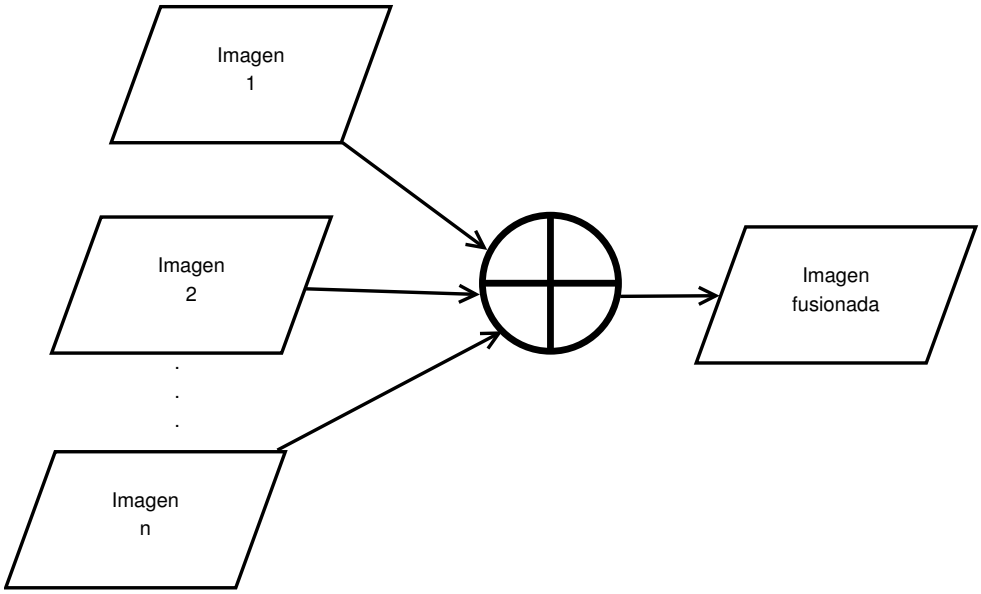


Figura 3.5: Diagrama de fusión simple

Algoritmo de fusión de exposición (EF)

Este algoritmo utiliza el método de fusión de exposición presentado en el marco teórico (sección 2.2.5), donde se obtiene una imagen fusionada a partir de diferentes imágenes de entrada con distintos niveles de intensidad y pesos para cada píxel.

Como las imágenes de geles utilizadas son adquiridas en la escala de grises se elimina el paramétro de saturación, puesto que éste solo afecta la calidad de la imagen si se utilizan imágenes a color (figura 3.6).

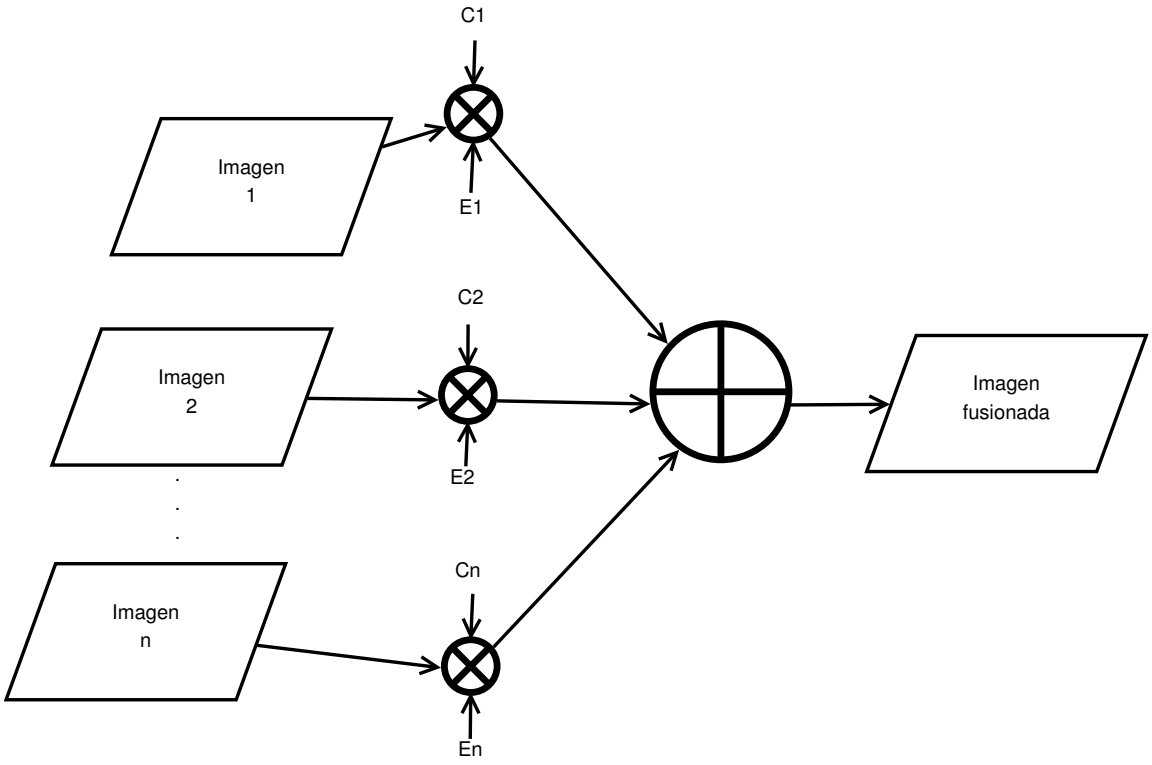


Figura 3.6: Diagrama de fusión de exposición

3.1.5 Sistema de adquisición y pre-procesamiento de mejoramiento de calidad de imágenes

El sistema final que se implementa para mejorar la calidad de las imágenes digitales de geles de electroforesis se muestra en la figura 3.7.

Como se observa en la figura 3.7 este sistema se puede dividir en:

1. Etapa de captura de imágenes
2. Etapa de fusión
3. Etapa de linealización

Etapa de captura de imágenes

En esta etapa se utiliza el sistema de adquisición optimizado de la cámara digital diseñado (sección 3.1.3) para capturar tres imágenes del gel de electroforesis en estudio a diferentes niveles de intensidad; baja intensidad luminosa (imagen oscura I_o), media intensidad luminosa (imagen media I_m) y alta intensidad luminosa (imagen clara I_c), donde cada imagen generada tiene características que la diferencian de las otras.

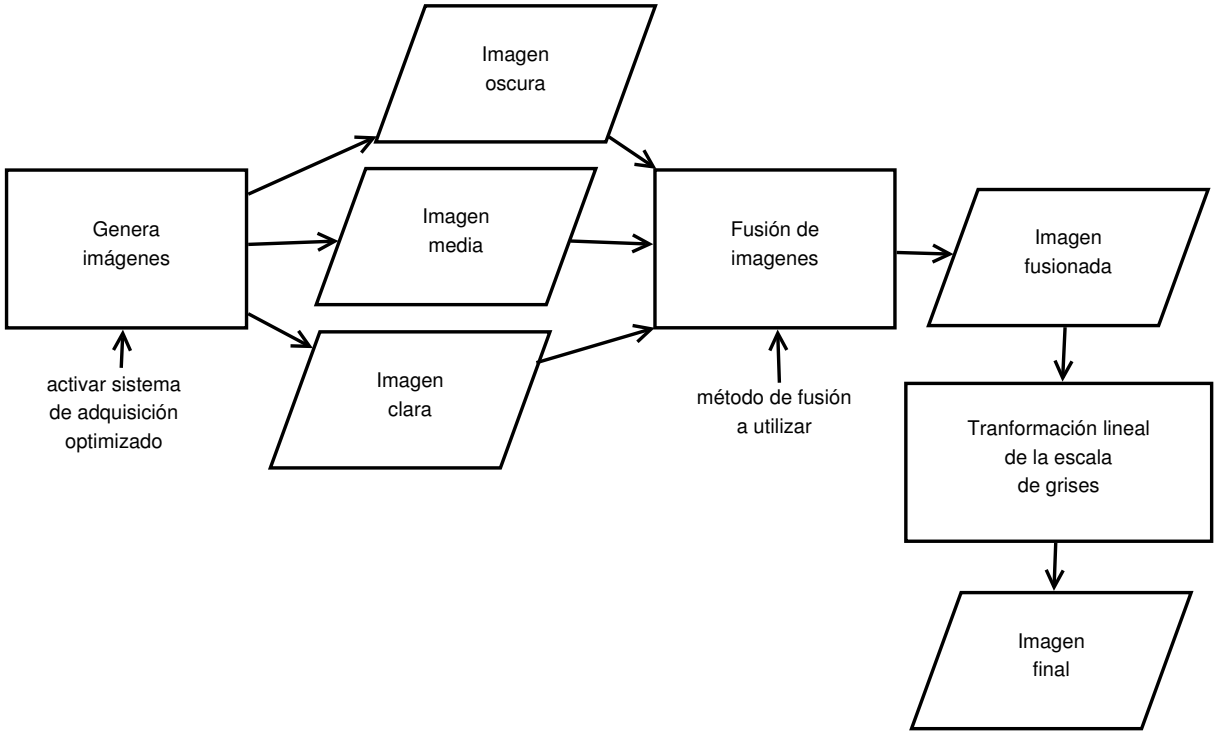


Figura 3.7: Diagrama de bloques del sistema de adquisición y pre-procesamiento de mejoramiento de calidad de imágenes

Etapa de fusión

En esta etapa se realiza el proceso de fusión de las imágenes generadas para obtener una imagen fusionada I_f con mejor calidad y mayor nivel de detalles. En el proceso de fusión se puede seleccionar cualquiera de los dos métodos implementados en este trabajo (sección 3.1.4).

Etapa de linealización

En esta etapa se pretende aumentar más los índices de calidad de la imagen fusionada mediante la transformación lineal de la escala de grises de ésta al intervalo $[0, 255]$, para obtener la imagen final I_{out} del sistema con la mejor calidad posible.

Transformación lineal de la escala de grises Este método de transformación convierte cada píxel de la imagen de entrada al intervalo $[p_{out_{min}}, p_{out_{max}}]$ deseado y consiste en aplicar la función lineal

$$f(p_{out}) = mp_{in} + b \tag{3.4}$$
$$m = \frac{p_{out_{max}} - p_{out_{min}}}{p_{in_{max}} - p_{in_{min}}}$$
$$b = p_{out_{max}} - mp_{in_{max}}$$

a cada uno de los píxeles de la imagen de entrada, donde $p_{in_{max}}$ y $p_{in_{min}}$ se obtienen de la imagen de entrada [32].

Para este trabajo se utiliza el intervalo máximo de linealización y (3.4) queda

$$\begin{aligned} f(p_{out}) &= mp_f + b \\ m &= \frac{255}{p_{f_{max}} - p_{f_{min}}} \\ b &= 255 - mp_{f_{max}} \end{aligned} \tag{3.5}$$

donde $p_{f_{max}}$ y $p_{f_{min}}$ se obtienen de la imagen fusionada.

3.2 Detección de carriles y rectificación de imágenes

Los métodos propuestos en el proyecto anterior [7] para la detección de carriles se basan en heurísticos adaptados a imágenes de AFLP que exhibiesen espacios entre carriles [28], con poca posibilidad de ser adaptados a otras configuraciones de imagen. Con la tesis de Pablo Barrantes [13, 14] desarrollada en el contexto de este proyecto, se exploraron métodos más robustos para dicha detección. La figura 3.8 presenta el diagrama de bloques del subsistema de detección de carriles y rectificación de la imagen para compensar las distorsiones que se presenten.

Para la detección de los carriles es necesaria una técnica de detección de bordes para la cual se exploraron en el marco del proyecto con la tesis de Pablo Barrantes [13] dos caminos: 1) utilizando redes neuronales 2) utilizando el gradiente. Además se utilizó en dicha tesis el operador de autocorrelación como un método para obtener el período de las columnas del gradiente, el cual representa el ancho de los carriles. La optimización mediante mínimos cuadrados fue utilizada para encontrar el polinomio que describe el comportamiento de los anchos de los carriles para distintas columnas del gradiente. También se utilizó un modelo matemático de distorsión óptica radial necesario para el entrenamiento de modelos activos de forma (ASM).

3.2.1 Creación del ASM

El método propuesto por Cootes et.al. [18] para la creación del modelo se basa en la ubicación manual de los puntos en el conjunto de entrenamiento; sin embargo, dado el costo en tiempo de esta labor y el hecho de que las distorsiones radiales están modeladas matemáticamente se crea el ASM de forma sintética, es decir, se basa en la aplicación de la distorsión radial a los puntos posicionados equidistantemente a lo largo del eje x de la imagen. Esto permite al usuario elegir el número de puntos que desea utilizar para el proceso. Así el ASM se crea de la siguiente forma:

1. Ubicación de la coordenada x para el número de puntos elegidos por el usuario.

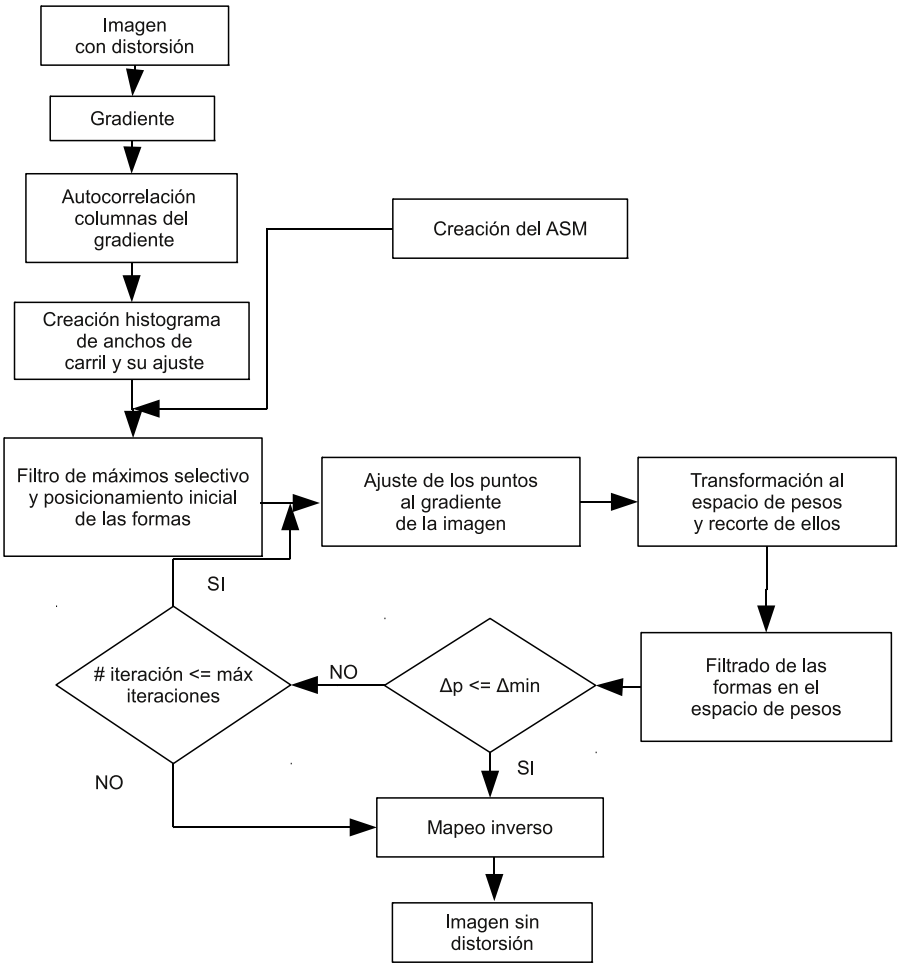


Figura 3.8: Diagrama de bloques de la detección de carriles y rectificación de la imagen.

Estos son equidistantes uno del otro, a una distancia S_p:

$$x_i = \left(k + \frac{1}{2}\right) S_p, \qquad k = 0, 1, \ldots, p-1$$

2. Aplicar el modelo de distorsión variando el centro de la imagen dentro del área indicada en la figura 3.9, en donde Δy es un parámetro elegido por el usuario y Δx es un porcentaje de Δy, elegido también por el usuario. Los parámetros a y d del modelo de distorsión

$$r_d(r_u) = r_u + ar_u^3 + br_u^5 + \cdots$$

son elegidos empíricamente eligiendo aquellos que se adaptan a las distorsiones radiales presentes en las imágenes de geles de electroforesis utilizadas.

3. Para cada variación del centro de la imagen considerar el conjunto de puntos distorsionados como una figura del conjunto de entrenamiento.

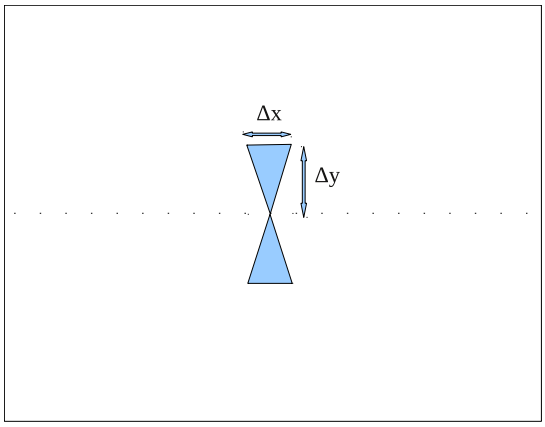


Figura 3.9: Desplazamiento del centreo de la imagen para generar la distorsión radial.

3.2.2 Gradiente para detección de bordes

Para la obtención del gradiente se parte de la imagen distorsionada en escala de grises. Debido a la aplicación, y asumiendo que los geles están alineados horizontalmente (carriles horizontales) la información relevante del gradiente es aquella a lo largo del eje y. Por lo tanto, se aplica el operador gradiente solamente en esta dimensión. Con el fin de intensificar la detección de bordes horizontales, se hace uso de un kernel con mayor número de columnas que de filas, como se muestra en la figura 3.10. Esta modificación en el

-1	-1	-1	-1	-1	-1
0	0	0	0	0	0
1	1	1	1	1	1

Figura 3.10: Kernel para detección de bordes horizontales.

número de columnas no es estática, si no que está dado por un porcentaje (definido por el usuario) de la separación entre los puntos S_p.

Los pasos para la obtención del gradiente son los siguientes:

1. Creación del kernel con el número de columnas deseadas.
2. Convolución del kernel con la imagen de entrada.
3. Si el gel tiene separación entre los carriles rectificar el gradiente. Esto es, para todo valor del gradiente $G(x, y)$ aplicar:

   $$\text{Si } G(x, y) < 0 \rightarrow G(x, y) = 0$$

   Esta operación se aplica ya que el gradiente en la dimensión y de la separación entre dos carriles tiene un perfil como el mostrado en la figura 3.11. Interesa conservar solamente uno de los bordes, por lo que al rectificar quedará solamente aquel en el que hay una transición entre carril (tonalidad oscura) y la separación (tonalidad clara).
4. Si el gel no tiene separación entre los carriles aplicar:

   $$G(x, y) \leftarrow |G(x, y)|$$

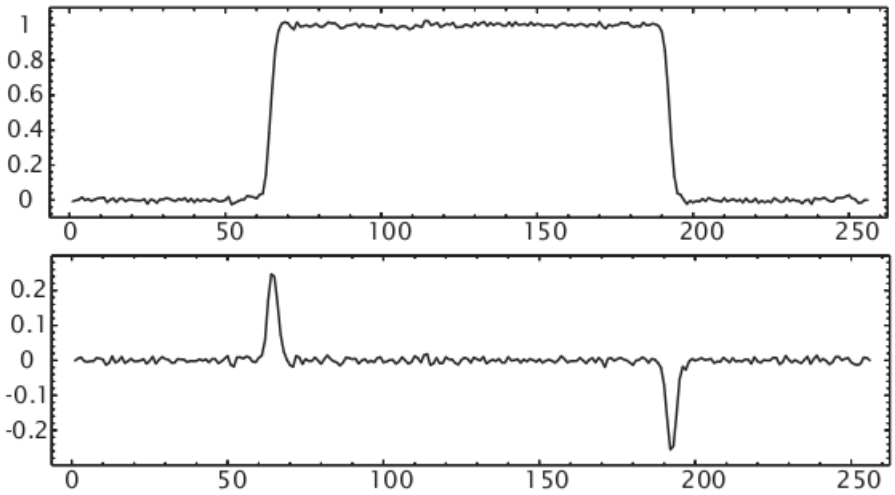


Figura 3.11: Primera derivada de una función unidimensional (tomado de [35])

Para este caso un cambio de tonalidad indica que existe un cambio de carril y no una separación por lo que se justifica la aplicación del valor absoluto.

5. Normalizar el gradiente.

3.2.3 Estimación del ancho de los carriles

El gradiente de la imagen será aproximadamente periódico en las áreas donde existen carriles, siendo el período el ancho de los carriles más la separación entre ellos, si ésta existe. Por lo tanto si se aplica la autocorrelación a una columna del gradiente se puede estimar el período de esa señal buscando el máximo más cercano al origen.

En este caso, se aplica la autocorrelación a p-columnas ($\underline{\mathbf{c}}$) para crear un histograma inicial de los anchos de carril. Con éste se obtiene el ancho promedio l_w presente en la imagen. Además se aplica un filtro guassiano a la autocorrelación para filtrar ruido presente en la imagen.

También se crea un vector $\underline{\mathbf{w}}$ que contiene los anchos de carril específicos para cada columna. Este se ajusta mediante un filtrado de mediana, partiendo de la premisa de que los cambios entre los anchos de los carriles no son pronunciados. Finalmente se calcula el polinomio de la forma $f(x) = ax^3 + bx^2 + cx + d$ que mejor se ajusta a la curva que describen los anchos de los carriles utilizando el método de los mínimos cuadrados [53] y se vuelve a ajustar el vector.

3.2.4 Detección de carriles

Se parte de cuatro premisas sobre los geles de electroforesis en los que se realizará la detección de carriles:

1. La orientación de los carriles es fija y conocida *a-priori* en las imágenes (horizontal o vertical).
2. Los carriles en un mismo gel tiene los mismos anchos y son separados con la misma distancia (que puede ser cero).
3. La magnitud máxima del gradiente de la imagen corresponde a los bordes del carril.
4. En la magnitud del gradiente en principio no deben ocurrir eventos visuales entre los carriles.

Partiendo de esta información se puede eliminar el ruido existente entre los carriles una vez aplicado el gradiente mediante un algoritmo de filtrado selectivo de máximos, resultando en una imagen que contiene información de dónde están ubicados los bordes de los carriles. Los pasos para este filtrado se detallan a continuación:

1. Aplicar a $\underline{\mathbf{c}}$ un filtro de máximos, utilizando para ello una máscara del tamaño indicado por el valor del vector $\underline{\mathbf{w}}$ correspondiente a cada columna c_i. Lo anterior seleccionará solo aquellos máximos que correspondan a bordes de carriles.
2. Con las columnas filtradas crear un nuevo histograma de anchos ($\underline{\mathbf{h}}$) de carril.
3. Utilizando $\underline{\mathbf{h}}$, eliminar de las columnas filtradas aquellos máximos que no posean vecinos a una distancia cercana a l_w. Esto se logra centrándose en un máximo m_i y a partir de este buscar aquellos cercanos a una distancia $l_w(m_{i-1}, m_{i+1})$ y ponderándolos con el valor de esta distancia en el histograma (figura 3.12). Si este valor es alto, el máximo mi tiene validez.

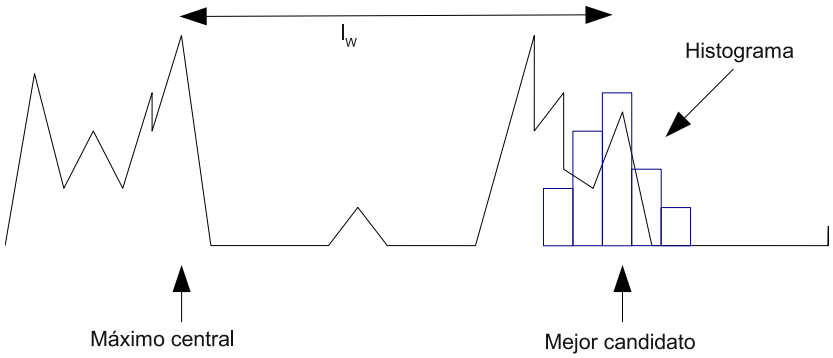


Figura 3.12: Búsqueda de máximos en las vecindades de otro máximo.

4. Hecho esto se procede a buscar los carriles más fuertes de la imagen. Estos serán el punto de partida para la detección de los demás. La técnica es utilizar la supresión no máxima del algoritmo de detección de bordes de Canny [16] en una imagen que esta formada a partir de $\underline{\mathbf{c}}$, la cual sería de tamaño $I_y \times p$ (figure 3.13). De esta forma se pueden encontrar las líneas de puntos conectados. Aquellas con mayor número de puntos serán los carriles más fuertes. Si no se encuentran líneas con un mínimo de puntos conectados (definido por el usuario) el paso 5 no se ejecuta y la entrada del paso 6 será el resultado del paso 3.
5. Partiendo de los carriles más fuertes, se procede de la misma forma que en el paso 3: Se buscan los máximos en la vecindad que son mejores candidatos a ser un borde de un carril y suprimiendo los valores intermedios (figura 3.14).

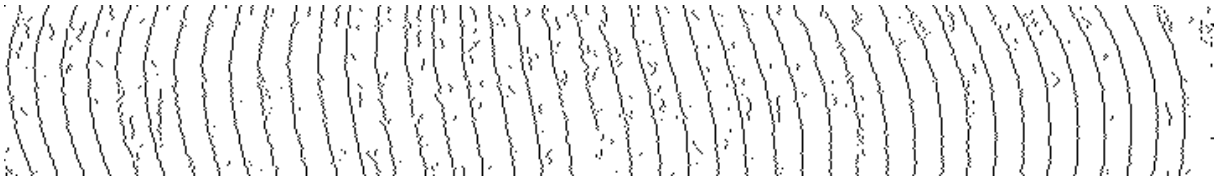


Figura 3.13: Ejemplo de supresión no máxima del algoritmo de Canny

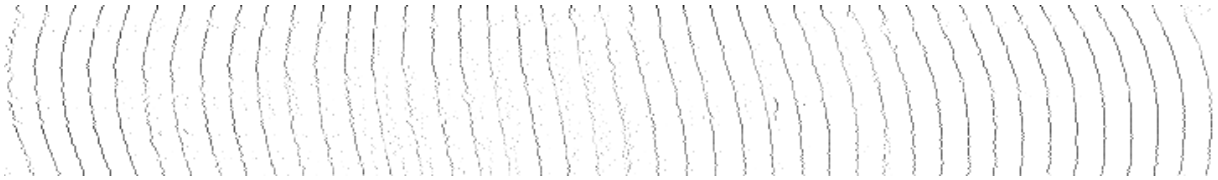


Figura 3.14: Supresión de máximos intermedios

6. Finalmente se aplica un filtro pasabajos a la imagen con el fin de rellenar espacios vacíos que serían potenciales candidatos a formar parte de un carril.

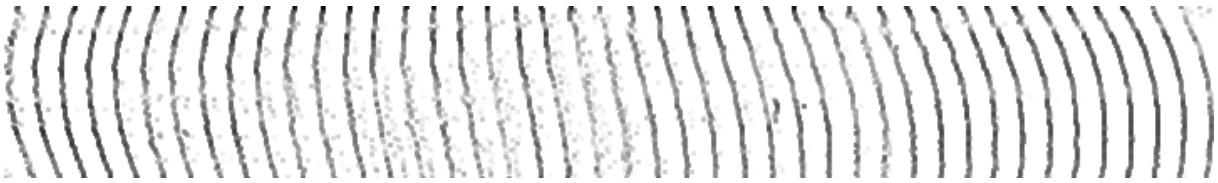


Figura 3.15: Resultado de aplicar un filtro pasabajos a la figura 3.14.

Siguiendo los pasos anteriores se obtiene una matriz o imagen (denotada con *maxFinal*) a partir de las p-columnas extraídas del gradiente, que contiene información de la ubicación de los carriles, la cual será el punto de partida para el posicionamiento inicial de las formas sobre la imagen.

3.2.5 Cálculo de la confiabilidad de una forma

La confiabilidad de una forma indica su validez o compatibilidad respecto a las otras formas. Para calcularla se evalúan todos sus puntos en *maxFinal* y se calcula el promedio. Para aquellos puntos que no correspondan exactamente a una fila de *maxFinal* se procede a hacer una interpolación lineal.

Sea una forma s conformada por un conjunto de puntos de la forma (x_z, y_z), la fila correspondiente para cada punto se obtiene como:

$$k_s = \frac{x_s}{S_p} - \frac{1}{2}$$

Así, siendo d_z el fraccionario de k_z y considerando *maxFinal* como una función $f(i, j)$, la confiabilidad de una forma está dada por

$$R_s = \sum_{s=0}^{p-1} f(y_z, k_s)(1 - d_z) + f(y_z, k_z + 1)d_z$$

3.2.6 Posicionamiento inicial de las formas

Las formas que se ubicarán sobre la imagen consisten de p-puntos que se encuentran separados por una distancia S_p, de forma que exista correspondencia con el entrenamiento del ASM. Estas representan tanto los bordes de los carriles como la distorsión que estos poseen. Se plantea una metodología similar a la de la sección anterior para ubicarlas.

1. Tomando una fila de *maxFinal* buscar el punto correspondiente al carril más fuerte encontrado en el punto 4 de la sección anterior y posicionar el primer punto de una forma, P_{10}.
2. Realizar una búsqueda hacia la izquierda del máximo más cercano a la distancia l y ubicar el primer punto de otra nueva forma, P_{20}.
3. Partiendo de P_{20} repetir el paso 2 para ubicar P_{30} y así sucesivamente hasta llegar al final de la imagen y ubicar P_{k0}.
4. Posicionándose de nuevo en el punto P_{10} realizar la búsqueda hacia la derecha para encontrar $P_{(k+1)0}$ y así sucesivamente hasta llegar al final de la imagen y posicionar el punto P_{n0}.
5. Repetir los pasos 1 al 4 hasta completar las filas, en otras palabras, llegar a los puntos P_{kp} y P_{np}
6. Calcular la confiabilidad de todas las formas y eliminar aquellas poco confiables, es decir, las que se ubicaron en secciones de la imagen donde no existen carriles.

3.2.7 Proceso iterativo de ajuste

Debido a que el posicionamiento inicial de las formas generalmente presenta imperfecciones, se requiere de un paso final previo a la rectificación de la imagen. Este se trata de un proceso iterativo en el cual las formas se van ajustando gradualmente a los bordes de los carriles, corrigiendo la posición de aquellos puntos que no se ajustan a las distorsiones para las cuales fue entrenado el ASM. El procedimiento termina cuando se alcanza una mínima variación entre iteraciones o cuando se alcanza un máximo número de iteraciones. Ambas condiciones son definidas por el usuario. Los detalles se presentan en [13].

3.2.8 Rectificación de la imagen

Una vez ajustadas las formas a los bordes de los carriles se tiene una cuadrícula no uniforme formada por puntos consecutivos. Por ejemplo, sea una forma $s_i(n)$ y otra $s_{i+1}(n)$ donde n representa los puntos, un elemento de la cuadrícula tendría como esquinas $(s_i(1), s_i(2), s_{i+1}(1), s_{i+1}(2))$. A partir de esta cuadrícula se crea una imagen nueva que convierta los cuadriláteros formados a rectángulos de largo S_p y ancho l utilizando un mapeo bilineal (detalles en [13]).

3.3 Estimación del efecto sonrisa

Uno de los efectos de mayor complejidad algoritmica en la automización de la detección automática de bandas y correspondencias es la detección del efecto sonrisa. La tesis de Antonio Aguilar [1] brindó resultados ya presentados en el informe del proyecto anterior [7] que se refinaron en este proyecto con la tesis de Pedro Alpízar [5, 4], y presentados a continuación.

El método propuesto se basa en obtener las líneas formadas por bandas correspondientes entre carriles en la imagen de electroforesis con la ayuda de la inclinación de las bandas. De aquí en adelante se les refiere a dichas líneas como líneas de bandas. Una vez creadas las líneas de bandas mediante el gradiente y algoritmos de detección de máximos se generan modelos de formas, los cuales son filtrados para eliminar todas aquellas formas cuyo aporte de información pueda entorpecer el proceso. Luego, mediante un proceso iterativo, las formas son modificadas para que el cambio de una a otra no sea abrupto y se apeguen a las deformaciones de la imagen. Con la información obtenida de los modelos de forma generados se realiza un mapeo inverso a la imagen para eliminar la distorsión efecto sonrisa.

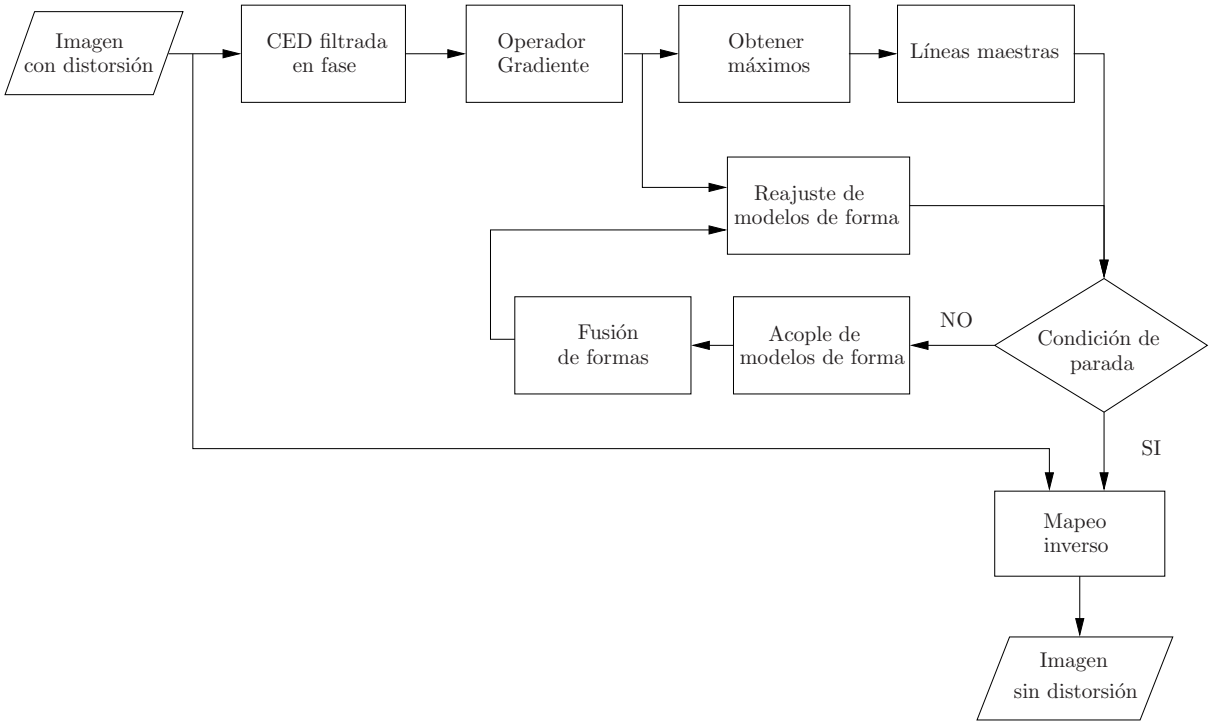


Figura 3.16: Solución propuesta para la corrección del efecto sonrisa

La propuesta anterior se ilustra en la figura 3.16, en donde el bloque "CED filtrada en fase" es el bloque encargado de crear las líneas de bandas. Los bloques "Operador gradiente", "Obtener máximos" y "Lineas maestras" son los encargados de obtener y filtrar los modelos de forma a utilizar. La condición de parada verifica que se cumpla un número de iteraciones dadas por el usuario o que los modelos de forma converjan, esto

quiere decir que no sean modificadas por los bloques "Fusión de formas" o "Reajuste de modelos de forma". Al final del proceso se realiza el mapeo inverso.

La transformación de las líneas de bandas a modelos de forma se realiza creando un vector que contenga la información de la desviación de los elementos con respecto a la media, ésta última también es guardada. Las formas tienen la misma extensión que la imagen en el lado paralelo a las líneas de bandas como se muestra en la figura 3.17, donde la linea negra continua representa el borde detectado de una línea de bandas, la línea punteada es el valor medio de la ubicación en x de los elementos que le pertenecen.

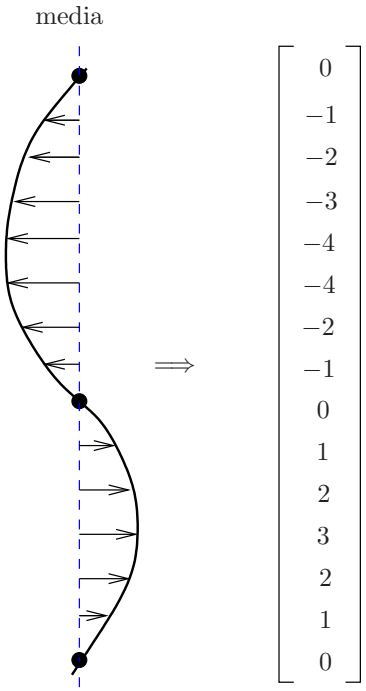


Figura 3.17: Trasformación de líneas a formas

Para explicar más en detalle los pasos de la propuesta se divide el sistema completo en cuatro etapas, mencionadas implícitamente en los párrafos anteriores. Estas etapas son:

- CED filtrada en fase.
- Detección de líneas de bandas.
- Ajuste de modelos de formas a las líneas de bandas.
- Corrección del efecto sonrisa.

3.3.1 CED filtrada en fase

Debido a que en la presente solución hay que leer la distorsión angular de las bandas para determinar las líneas de bandas de la imagen se decide utilizar una dispersión anisotrópica que permita generar dichas líneas, lo cual evita utilizar algoritmos que tomen la dirección de las bandas e interpolen cuáles son correspondientes y cuáles no. Debido a las líneas de bandas no tienen cambios bruscos, estas pueden ser vistas como flujo en la imagen se

utiliza CED.

Utilizar CED permite seguir el flujo natural de las imágenes. En las imágenes de electroforesis el flujo predominante está dado en la dirección de los carriles, ya que estos se encuentran fuertemente definidos en la imagen. Esta condición impide que se pueda dar una aplicación directa de CED sobre la imagen de electroforesis.

Para obtener una difusión con la información de las deformaciones en las bandas y producir líneas de bandas, ignorando la información provista por los carriles de la imagen se cambia la manera en que se obtiene coeficiente estructural de la difusión. El cambio se realiza para que el sistema utilice

$$\hat{J}_\rho = K_\rho * \left(\hat{\nabla} f_\sigma \hat{\nabla} f_\sigma^T\right) \tag{3.6}$$

donde

$$\hat{\nabla} f_\sigma = |\nabla f_\sigma| \cdot p(\arg \nabla f_\sigma) \cdot e_{\arg \nabla f_\sigma} \tag{3.7}$$

en lugar de (2.18). El resultado obtenido es una CED filtrada en fase. Se le denomina filtrada en fase ya que afecta la magnitud del gradiente de la imagen dependiendo de su fase. Como se observa en (3.7), la función p es la encargada de realizar esta labor, mientras que el vector unitario $e_{\arg \nabla f_\sigma}$ conserva la dirección del gradiente original para no perder información de la dirección. La función p debe cumplir con las siguientes características:

- Simétrica.
- Centrada en dirección a los carriles de la imagen.
- Periodicidad π.

La simetría es para evitar que el flujo se vea afectado por el filtro y se obtengan falsas tendencias en la imagen. El ángulo de centro debe ser el mismo que el ángulo del flujo los carriles ya que el coeficiente estructural utiliza el gradiente como fuente de información y como se explicó anteriormente el gradiente y el flujo tienen una diferencia de 90º. De esta forma si se coloca el filtro con el flujo de los carriles se deja intacta la información del gradiente de las bandas el cual está a 90º con relación a los carriles. La periodicidad es necesaria para evitar filtrar los componentes negativos del gradiente.

Para la presente solución se propone p como

$$p(\theta) = \begin{cases} \cos\left(\frac{\pi(\theta+\phi-\varphi-\gamma-\pi)}{2\cdot\gamma}\right) & \text{si } \theta \in [\phi-\varphi-\gamma, \phi-\varphi] \\ 1 & \text{si } \theta \in [\phi-\varphi, \phi+\varphi] \\ \cos\left(\frac{\pi(\theta+\phi+\varphi)}{2\cdot\gamma}\right) & \text{si } \theta \in [\phi+\varphi, \phi+\varphi+\gamma] \\ p(\theta+k\pi) & \text{con } k \in \mathbb{Z} \\ 0 & \text{el resto} \end{cases} \tag{3.8}$$

Donde θ es un la variable independiente angular, $2 \cdot \varphi$ es el ancho del segmento que se conserva constante, γ es el ancho de las componentes cosenoidales del filtro, esto se puede observar en la figura 3.18 donde se muestra la función $p(\theta)$ para un $\theta \in [0, 2\pi]$, un $\phi = 0$ y $\varphi = \gamma = \frac{\pi}{8}$.

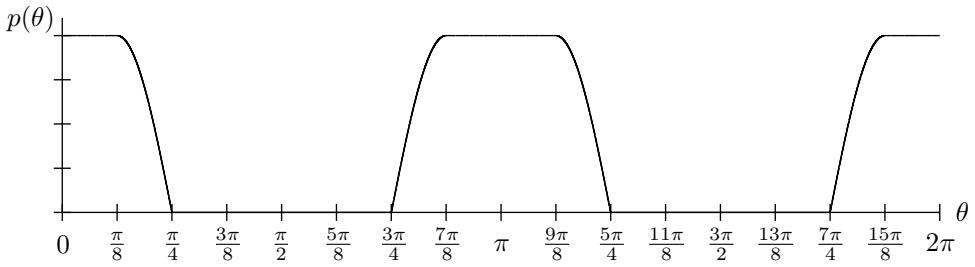
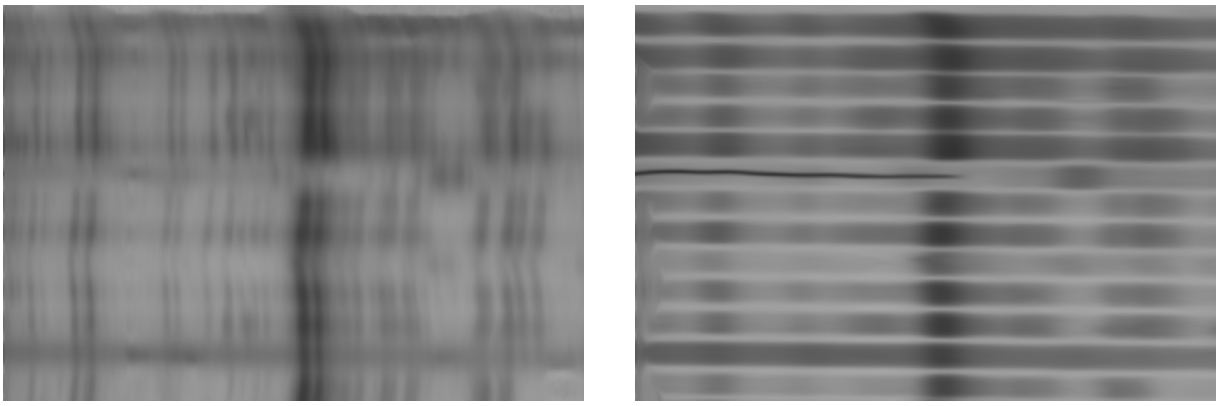


Figura 3.18: Filtro de fase p con $\phi = 0$ y $\varphi = \gamma = \frac{\pi}{8}$

En la figura 3.19 se pueden observar las diferencias entre aplicar a una imagen de electroforesis la CED filtrada en fase, figura 3.19(a), y aplicar CED como esta definida en [64], figura 3.19(b). En la primera de las imágenes se obtiene una difusión que forma líneas de bandas (lo buscado) mientras que en la segunda se obtiene una difusión a lo largo de los carriles.



(a) (b)

Figura 3.19: Imagen comparativa entre una imagen utilizando CED filtrada (a) y CED sin filtrar (b)

3.3.2 Detección de líneas de bandas

Luego de aplicar CED a la imagen del proceso de electroforesis se obtienen sus líneas de bandas. Ellas describen el efecto sonrisa de la misma. Para leer la información de dichas bandas se utiliza el operador gradiente, utilizando la máscara tipo Sobel.

El gradiente de una imagen lee los bordes presentes en ella, en este caso dos bordes por cada línea de bandas, lo cual es indeseable ya que produce exceso de información sobre la distorsión de la imagen. Para eliminar el exceso de información se filtran los datos del gradiente suprimiendo los valores negativos del gradiente, asegurando de esta manera la lectura de sólo uno de los bordes de cada línea de bandas.

Ahora los bordes de líneas de bandas deben ser convertidos a modelos de forma que describan la distorsión presente en la imagen. Dado que las imágenes poseen líneas de

bandas de diferentes tamaños y ubicadas a distancias no homegeneas entre sí se deben filtrar las formas que van a ser utilizadas con la finalidad de evitar cruces entre ellas o crear formas que no tengan suficiente respaldo en la imagen y produzcan distorsión en lugar de corregir la imagen. Los criterios utilizados para el filtro son la longitud de la línea de bandas y la intensidad del borde de la banda. Las líneas seleccionadas son las líneas de bandas principales o maestras de la imagen.

El primer paso en la creación de los modelos de forma es filtrar los bordes de líneas de bandas según la intensidad del borde. Este filtro se aplica mediante un filtro de máximos y un algoritmo comparador. Con el filtro de máximos se rellena un área determinada con el valor máximo de esa área. Esa región queda delimitada por un parámetro definido por el usuario. Una vez obtenida la imagen filtrada en máximos se utiliza el comparador para extraer únicamente los elementos máximos de ella. El resultado es una imagen que tiene líneas de bandas de un píxel de grosor aproximadamente.

Con la ayuda de supresión de no máximos se inserta la imagen filtrada en máximos para eliminar el valor del píxel y asegurar que los puntos máximos detectados por el filtro de máximos pertenezcan a bordes de líneas de bandas. Una vez eliminada la información del valor del píxel se agrupan los píxeles que están unidos unos a otros, obtenido de ésta líneas de bandas separadas en grupos. En este punto se aplica el segundo criterio de filtro: longitud de las formas. Este filtro solo dejará pasar aquellos grupos que contengan un mínimo de elementos, definido por el usuario.

Con los grupos filtrados se generan los modelos de forma que representarán la distorsión del efecto sonrisa en la imagen. A cada uno de los grupos se le calcula el valor medio y se ordenan según éste. Los modelos de forma son colocados en una matriz y un vector. La matriz almacena la forma de la manera descrita al inicio del presente capítulo, una forma por columna. Dicha matriz se denomina $\mathbf{D}$. El vector denominado $\underline{\mathbf{m}}$ tiene como finalidad almacenar los valores medios de cada una de las formas.

Además de la matriz $\mathbf{D}$ se crea la matriz $\mathbf{C}$ con la confianza de cada uno de los elementos de las formas. Ésta confianza es dada por la magnitud del gradiente en la ubicación respectiva de cada elemento.

Las formas en este momento solo contienen información en aquellos lugares donde los grupos tenían información. El resto de los elementos tiene un valor constante, cero para la matriz de desplazamientos ($\mathbf{D}$) y un valor mínimo ε en la matriz de confianza($\mathbf{C}$).

3.3.3 Ajuste de modelos de formas a las líneas de bandas

La etapa de ajuste de los modelos de forma a las líneas de bandas busca que los modelos de forma dejen de ser fragmentos de información y se conviertan en formas que describan fielmente el efecto sonrisa presente en toda la imagen. Para ello el proceso de ajuste de las formas se realiza en cuatro pasos:

- Difundir con confiabilidad.

- Fusionar formas.
- Reubicar las formas.
- Redefinir la estructura de las formas.

Estos cuatro pasos se realizan de manera cíclica hasta cumplir con la condición de parada (se cumpla una cantidad de iteraciones o no exista cambio en las formas). En la figura 3.20 se muestra el ciclo de ajuste de las formas, en donde el bloque "Reajuste de modelos de forma" unifica los pasos de reubicación y redefinición antes mencionados.

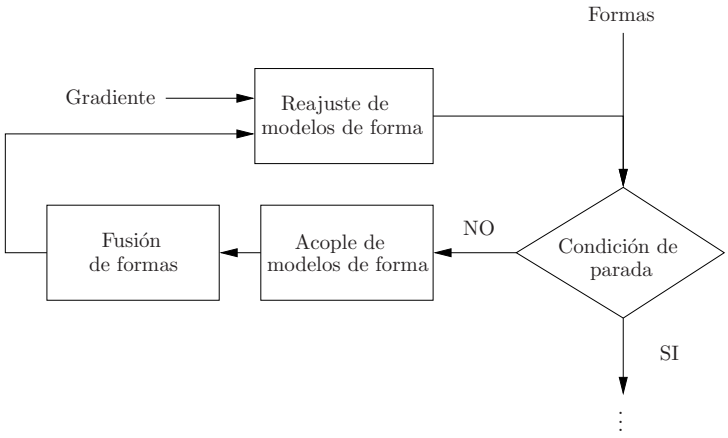


Figura 3.20: Ciclo para el acople de los modelos de formas a la distorsión de la imagen

En el bloque de "Acople de modelos de forma" se implementa la dispersión con confianza explicada en la sección 2.4.2. Este bloque permite utilizar la información de formas adyacentes para completar aquellos sectores de ellas que no tienen respaldo de la imagen o es muy pobre. Además, éste bloque logra una transición suave entre las formas, como lo son los cambios en la curvatura del efecto sonrisa.

Se realiza seguidamente la "Fusión de formas". Esta fusión tiene como tarea unir dos formas adyacentes que representa una misma línea de bandas en la imagen. Para determinar si representan dos partes de una misma línea de bandas primero se considera que la diferencia de las medias de ambas formas se encuentre dentro de un rango definido por el usuario. De encontrarse en dicho rango se comprueba si pertenecen a dos partes de la misma línea de bandas (la confianza de sus elementos no se sobreponen) o a dos líneas de bandas independientes muy cercanas (la confianza de un porcentaje dado de sus elementos se sobreponen). En el caso en que se puedan fusionar, se eliminan las formas padres y se inserta la nueva forma que es la combinación de las dos anteriores, esto se realiza tanto en **D** como en **C**.

La dispersión con confiabilidad puede cambiar la estructura de las formas, desplazando los elementos de ellas fuera de los bordes de líneas de bandas. Para devolver los modelos de forma a los bordes de líneas de bandas se implementa el bloque "Reajuste de modelos de forma" el cual varía primero la ubicación total de la forma en la imagen, para encontrar el punto en que tiene un mayor acople. El acople de la forma se mide mediante la suma

de las multiplicaciones de la confianza de cada elemento por el valor del gradiente en el nuevo punto. Se resume como

$$acople = \sum_{j=0}^{n-1} \mathbf{C}_{i,j} \cdot \nabla f(\underline{\mathbf{m}}_i + \mathbf{D}_{i,j}, j) \tag{3.9}$$

donde n es la cantidad de elementos que posee la forma, i es el índice que denota la forma que se está analizando, f es la imagen analizada.

Luego de realizar el análisis de ubicación se realiza la redefinición de la estructura misma de la forma. Esto asegura que las formas mantengan sus elementos siempre en el valor máximo del borde de las líneas de bandas modeladas. La redefinición de la forma se lleva acabo reposicionando cada elemento de la forma al punto donde exista una mayor intensidad del gradiente que en el punto actual dentro de un número determinado de píxeles cercanos.

3.3.4 Corrección del efecto sonrisa

Las formas en este punto ya describen la distorsión presente en toda la imagen no solo por fragmentos como fue leída inicialmente. Con el modelo de formas se busca mapear la imagen hacia otra sin distorsión. Dado que las formas cubren la imagen de lado a lado en la dirección de las líneas de bandas, se puede realizar el mapeo entre las formas píxel a píxel mediante un mapeo lineal.

El mapeo lineal lo que busca es mapear cada píxel entre dos formas hacia la nueva región en la imagen corregida que se encuentra delimitada por los dos puntos medios de dichas formas como se muestra en la Figura 3.21.

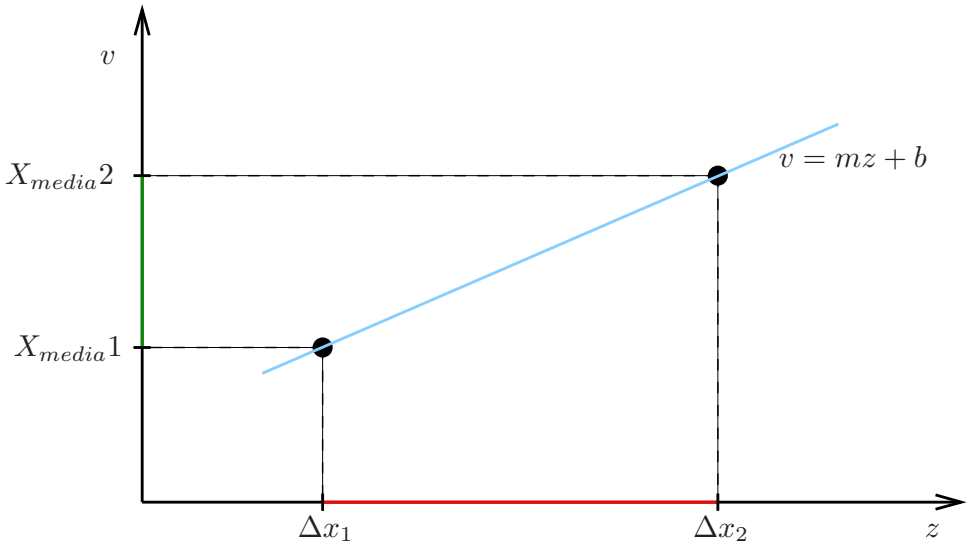


Figura 3.21: Mapeo inverso para corregir el efecto sonrisa

En la figura 3.21 la región roja entre Δx_1 (posición del elemento de la forma izquierda) hasta Δx_2 (posición del mismo elemento de la forma derecha), es convertida con ayuda

de

$$v = mz + b \tag{3.10}$$

donde las constantes m y b de se obtienen de resolver el sistema de ecuaciones

$$\begin{aligned} X_{media}1 &= m\Delta x_1 + b \\ X_{media}2 &= m\Delta x_2 + b \end{aligned} \tag{3.11}$$

en la región verde, que se extiende desde $X_{media}1$ (valor medio de la forma izquierda) hasta $X_{media}2$ (valor medio de la forma derecha). Esto debe ser realizado para cada pareja de elementos entre dos formas y para cada par de formas en la imagen, tomando los bordes como formas con desplazamiento cero y media su valor de posición.

El mapeo se realiza de los píxeles de la imagen corregida hacia los píxeles de la imagen con distorsión, para evitar un sobremuestreo o un submuestreo de la imagen corregida. El resultado del mapeo es una imagen sin distorsión de efecto sonrisa.

3.4 Detección de bandas

Todo el análisis de imágenes de geles de electroforesis se basa en la precisa detección de las bandas en los carriles. Tres tesis en el marco del proyecto se concentraron en esta tarea. En la tesis de David Soto [59, 58] se planteó el problema como un problema de optimización, que requiere parámetros de entrada como el ancho de las bandas, los cuales no siempre están disponibles. El trabajo de Edison Fernández [26] exploró el uso de espacios de escala para la determinación automática de dicho ancho. En la tesis de Randall Esquivel [22] se realizó una nueva propuesta que integra los resultados de las tesis de Soto y Fernández.

3.4.1 Detección por medio de optimización

La figura 3.22 muestra el diagrama de flujo de la solución propuesta para la detección de bandas planteado como problema de optimización. Ésta se basa en el análisis individual de cada carril horizontal presente en la imagen de geles, específicamente en la distribución de intensidad del carril. El método es capaz de realizar un análisis basado en la proyección promedio de las cinco filas centrales del carril sobre un vector, o de analizar en forma paralela la distribución de intensidad de una cantidad de filas establecidas por el usuario. Estas filas son seleccionadas de la fila central hacia los extremos del carril. De este modo se reduce la influencia del ruido en el análisis.

La estrategia se fundamenta en la forma gaussiana típica descrita por el perfil de intensidad de las bandas [44], tal que, la distribución de intensidad presente a lo largo de un carril se puede modelar como una sumatoria de funciones gaussianas, cuyos parámetros describen a cada una de las bandas, esto eliminando previamente lo considerado como fondo de la imagen del carril (un algoritmo para esto fue propuesto en el proyecto anterior [7]). En

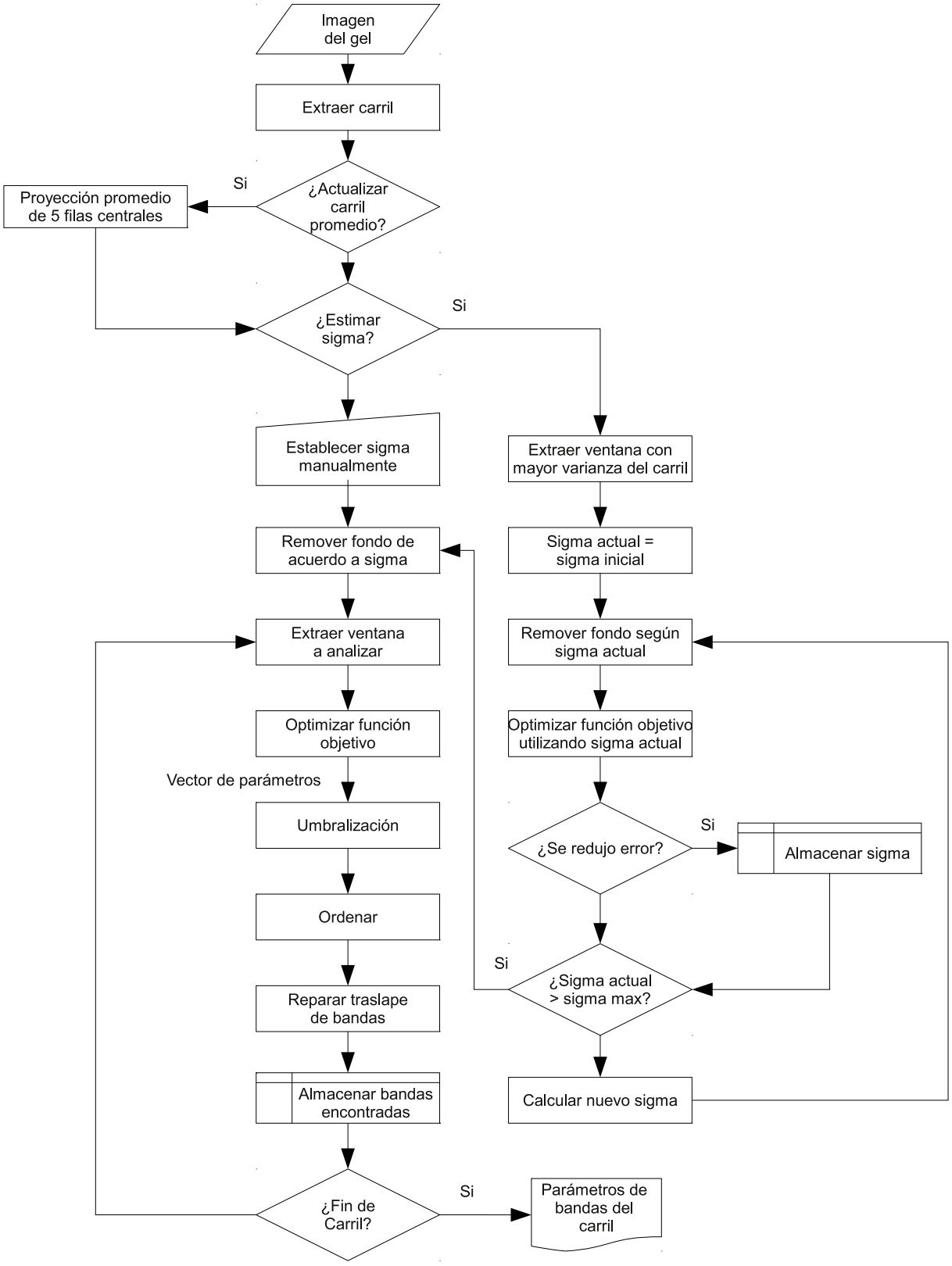


Figura 3.22: Diagrama de flujo del sistema propuesto de detección de bandas con un optimizador.

la solución propuesta se adopta como criterio de diseño que todas las bandas presentes en el carril tienen una misma varianza σ^2, la cual es un parámetro de entrada para el algoritmo. Los restantes parámetros, amplitud y valor medio, son estimados mediante el uso de un método de optimización de parámetros en conjunto con el algoritmo genético PESA [19] y los frentes de Pareto, utilizando como función objetivo una función de error cuadrático medio entre la distribución de intensidad del carril y la sumatoria de funciones gaussianas.

A partir de una imagen rectificada, se extrae la imagen del carril de dimensiones $n \times m$ con m filas y n columnas, se realiza un análisis de la distribución de intensidad. Este análisis puede ser realizado de dos formas diferentes. La primera de ellas considerando de forma paralela la distribución de intensidad de p filas a partir de la fila central de la imagen del carril, siendo p un parámetro establecido manualmente. Este tipo de análisis da como resultado p vectores $\underline{\mathbf{v}}_k$ con $k \in \{0, 1, 2, \ldots, p-1\}$, cada uno de dimensión N. Por otra parte, cada banda encontrada b_u es descrita por los parámetros de amplitud A_u y media μ_u de su respectiva función gaussiana. Así para el caso de la banda b_0, sus parámetros A_0 y μ_0 en cada una de las filas en consideración se encontrarán respectivamente en la primera y segunda dimensión de cada vector $\underline{\mathbf{v}}_k$. Para este caso el algoritmo da como resultado que los parámetros de las bandas ubicadas dentro del carril corresponden a la mediana M_d de cada una de las N dimensiones de los p vectores.

Lo anterior puede ser expresado en forma matricial, considerando una matriz $\mathbf{Y}$ de dimensión $p \times N$, $(N = 2N_b)$ formada por cada uno de los p vectores. El vector resultante $\underline{\mathbf{v}}_f$ que contiene los parámetros que describen a las bandas presentes en el carril se obtiene de la mediana de cada una de las columnas de $\mathbf{Y}$.

$$\mathbf{Y} = \begin{bmatrix} A_0^{(0)} & \mu_0^{(0)} & \cdots & A_{N_b}^{(0)} & \mu_{N_b}^{(0)} \\ A_0^{(1)} & \mu_0^{(1)} & \cdots & A_{N_b}^{(1)} & \mu_{N_b}^{(1)} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ A_0^{(p-1)} & \mu_0^{(p-1)} & \cdots & A_{N_b}^{(p-1)} & \mu_{N_b}^{(p-1)} \end{bmatrix}$$

$$\underline{\mathbf{v}}_f = \begin{bmatrix} M_d(\mathbf{Y}_{C_0}) & M_d(\mathbf{Y}_{C_1}) & \cdots & M_d(\mathbf{Y}_{C_{N-2}}) & M_d(\mathbf{Y}_{C_{N-1}}) \end{bmatrix}^T$$

La segunda forma de análisis consiste en realizar una proyección sobre la fila central de la imagen del carril, utilizando ésta como vector de intensidades de referencia para la función objetivo a optimizar. Esta proyección consiste en asignar a la posición i, con $i \in \{0, 1, \ldots, n-1\}$ del vector central el promedio de los valores de intensidad del elemento i de las cinco filas centrales de la imagen del carril. De esta forma se disminuye el tiempo de procesamiento del primer método y se considera la información presente en las filas vecinas de la fila central. Lo anterior se expresa matemáticamente para cada posición i como:

$$D_i = \frac{1}{5} \sum_{j=j_c-2}^{j_c+2} D_{j,i}$$

donde $D_{j,i}$ corresponde al valor de la fila j y columna i de la imagen del carril y j_c la posición de la fila central.

Segmentación del carril en ventanas

El carril con el fondo previamente eliminado es segmentado en h ventanas de tamaño finito $w \times n$ con w un parámetro del método. Esto permite disminuir la cantidad de iteraciones necesarias por el algoritmo de optimización de parámetros Downhill Simplex [53] para realizar el análisis del carril completo, ya que la cantidad de iteraciones totales i_t necesarias para converger al mínimo de la función objetivo está en función de la cantidad de dimensiones del problema de optimización y de la superficie de error descrita por la función objetivo, y para el peor de los casos esta función es exponencial.

Así dividiendo el análisis total del carril en ventanas y limitando la cantidad máxima de bandas por ventana (B_{pv}) se logra acelerar la ubicación de las bandas presentes a lo largo todo del carril. Esta segmentación se ilustra en la figura 3.23, donde el traslape existente entre dos ventanas contiguas es de 8σ, distancia suficiente para abarcar la distribución total de una banda.

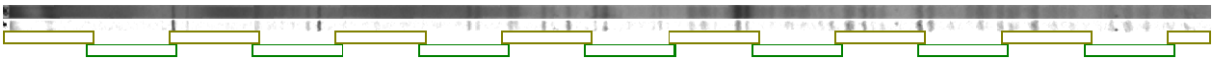


Figura 3.23: Recorrido del carril con ventanas en cascada.

Optimización de la función objetivo

La función objetivo a optimizar para una ventana de tamaño $w \times m$ es la función de error cuadrático medio (ECM) entre una de las filas del segmento de carril encerrado en la ventana y una sumatoria de funciones gaussianas que representan a cada posible banda dentro del segmento analizado:

$$ECM = \frac{1}{w}\sum_{i=0}^{w-1}\left[D(i) - \sum_{u=0}^{B_{pv}-1} A_u e^{-\frac{(i-\mu_u)^2}{2\sigma^2}}\right]^2 \tag{3.12}$$

donde $D(i)$ es el valor i de la fila analizada y el parámetro A_u el nivel de intensidad de la banda u en su píxel central μ_u siendo estos dos últimos los parámetros optimizados por el algoritmo, ya que σ es determinada en las etapas previas.

Para restringir el proceso de optimización a únicamente valores válidos, los parámetros μ_u y A_u se sustituyen por funciones sigmoides de un parámetro indirecto. Los detalles se presentan en [59].

La función ECM se puede interpretar como una superficie de error ubicada en un espacio de parámetros $setP^N$ con $N = 2B_{pv}$. Esta superficie de error contiene mínimos locales y puntos de silla, por lo cual no es posible asegurar la convergencia de un algoritmo optimizador al mínimo global de la función partiendo de cualquier punto inicial. La presencia de puntos de silla en la función provoca que el uso de algoritmos de optimización basados en la información brindada por la derivada de la función, como el método de gradientes conjugados, no sea una solución al presente problema de optimización. Por

este motivo en este trabajo se utiliza el algoritmo de optimización Downhill Simplex, el cual de igual forma no asegura la convergencia al mínimo global de superficies de error con mínimos locales, sin embargo, no es sensible a los puntos de silla ya que la optimización se basa únicamente en la información brindada por la función objetivo.

La solución presentada propone optimizar la función objetivo a partir de diferentes puntos ubicados en distintas regiones sobre la superficie de error y así facilitar la búsqueda del mínimo global de la función ECM. Se propone como solución el uso de los algoritmos genéticos como generadores de puntos N-dimensionales a optimizar, en combinación con el optimizador Downhill Simplex para asegurar la ubicación del mínimo de la función ECM o de al menos un punto que aproxime la distribución de intensidad de la ventana evaluada. El proceso completo se detalla en [59, 58].

3.4.2 Incorporación de estimación del ancho de las bandas

La detección de cantidad, posición, amplitud y ancho de las bandas en un carril se concretó con la tesis de Randall Esquivel [22], en la que se incorporan resultados de la tesis de Edison Fernández [26] sobre el uso del espacio de escalas, y las estrategias de aceleración y optimización algorítmica de David Soto [59]. El esquema general de dicha propuesta se ilustra en la figura 3.24.

La primera aproximación de la desviación estándar se obtiene realizando el cálculo del espacio de escalas, el cual permite ubicar los puntos máximos correspondientes a cada una de las bandas y con ello determinar la desviación (σ) inicial. El valor de la desviación estándar obtenido es luego utilizado para realizar el filtrado y segunda derivada del carril dependiendo del valor de σ obtenido. Los valores máximos detectados con la segunda derivada permiten obtener la posición de las bandas, con lo cual puede realizarse una segunda aproximación refinada de la desviación estándar ignorando los máximos cuya posición no corresponden a los máximos de la segunda derivada.

El valor de desviación estándar obtenido se utiliza en el siguiente paso de filtrado de la señal para eliminar el fondo y ruido indeseado, ya que este método necesita conocer el ancho de las bandas a optimizar para no deformar los datos. Cuando se obtienen el carril filtrado se hace una aproximación de σ para las posiciones encontradas con la segunda derivada utilizando minimización lineal para obtener el valor más probable de desviación estándar de las bandas. Este proceso se hace en forma iterativa hasta que la diferencia entre σ actual y el anterior sea menor a una tolerancia e establecida de 0.001.

Finalmente el algoritmo realiza una optimización multidimensional de la amplitud y posición de las bandas con el valor de σ obtenido, para lo cual se utiliza el algoritmo *Downhill Simplex* [53].

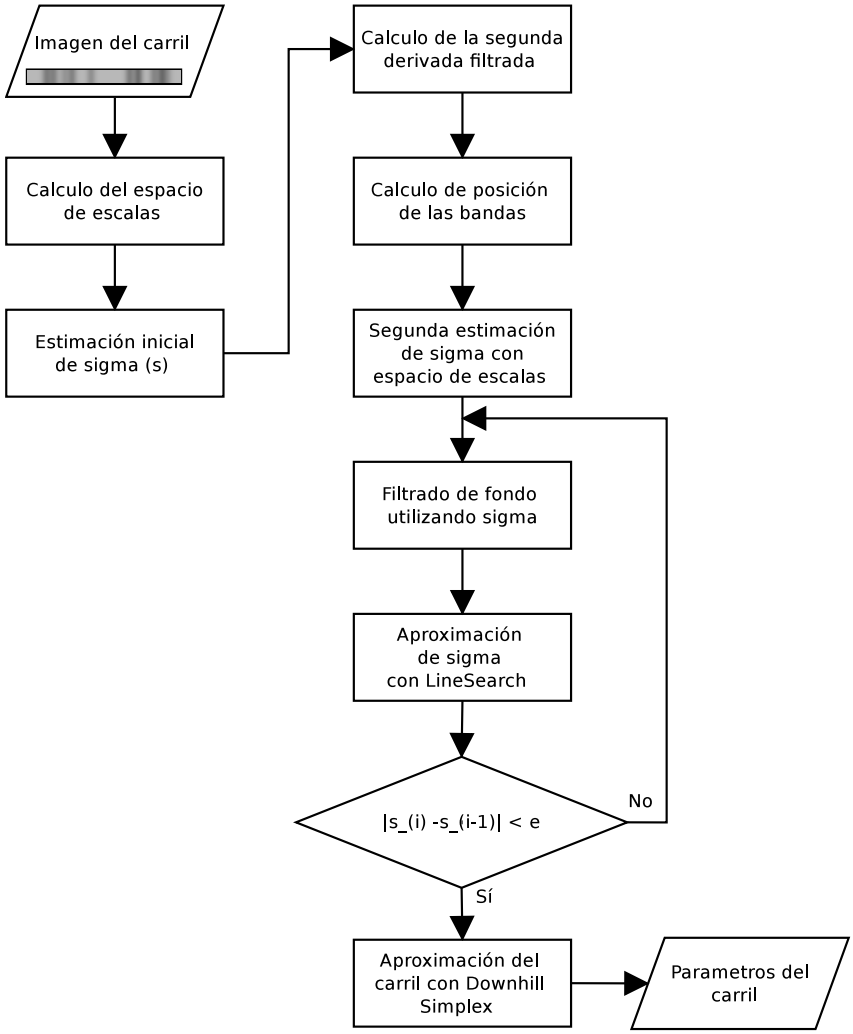


Figura 3.24: Diagrama completo del sistema propuesto.

Primera estimación del ancho de las bandas

Considerando que los carriles de geles de electroforessis a analizar están compuestos por bandas con distribución gaussiana [44], la construcción del espacio de escalas [39] involucra la convolución de dos funciones gaussianas de diferente desviación estándar. Sea $f(x)$ el perfil de intensidad de la banda y $g(x)$ la respuesta al impulso del filtro gaussiano para la escala σ_f:

$$f(x;\sigma_b) = \frac{1}{\sigma_b\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x}{\sigma_b}\right)^2}$$
$$g(x;\sigma_f) = \frac{1}{\sigma_f\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x}{\sigma_f}\right)^2}$$

donde σ_b representa la desviación estándar de la banda en el carril de electroforesis y σ_f corresponde a la desviación estándar del filtro aplicado. En [26] se presenta la deducción completa de las fórmulas utilizadas en esta sección. La convolución de las dos funciones

anteriores genera una nueva función gaussiana $h(x)$

$$h(x;\sigma_r) = \frac{1}{\sigma_r\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x}{\sigma_r}\right)^2} \tag{3.13}$$

donde

$$\sigma_r = \sqrt{{\sigma_b}^2 + {\sigma_f}^2} \tag{3.14}$$

Al aplicar la segunda derivada a una banda con centro en $\mu = 0$, por simplicidad, se obtiene:

$$\frac{\partial^2 h(x;\sigma_r)}{\partial^2 x} = h(x;\sigma_r)\frac{x^2 - \sigma_r^2}{\sigma_r^4} \tag{3.15}$$

En el caso de una única banda, el máximo ocurre en su centro (figura 3.25), esto es, en

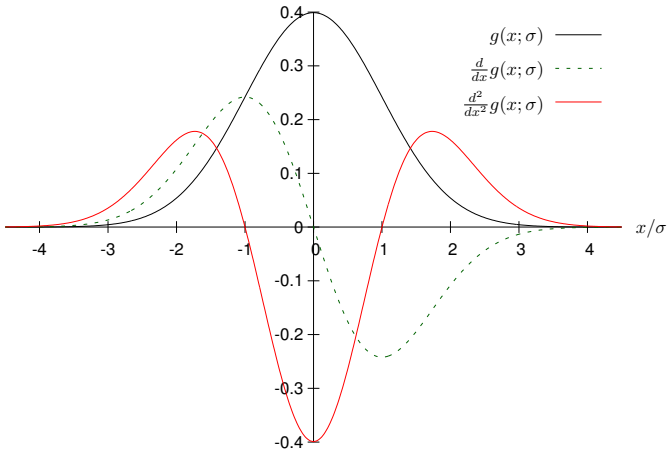


Figura 3.25: Respuesta gaussiana y sus primeras dos derivadas.

$x = \mu$, por lo que la expresión anterior puede simplificarse para obtener

$$\left.\frac{\partial^2 h(x;\sigma_r)}{\partial^2 x}\right|_{x=0} = \frac{1}{\sqrt{2\pi}\left(\sigma_b^2 + \sigma_f^2\right)^{3/2}} \tag{3.16}$$

Analizando esta ecuación se observa que el punto máximo depende del valor de la desviación estándar de la banda tanto como de la del filtro. La figura 3.26 muestra el comportamiento de esta función cuando se le normaliza multiplicando por la desviación estándar del filtro a diferentes potencias. Para la normalización con $\sigma_f^{3/2}$ se observa que el máximo del espacio de escalas coincide con la desviación estándar de la banda. Por lo tanto el valor 3/2 debe usarse para obtener la desviación estándar de la banda, y por esta razón es el valor utilizado en el programa.

Enventanado del espacio de escalas

El espacio de escalas normalizado presenta una forma directa para la medición de la desviación estándar de las funciones gaussianas en un carril. En general las imágenes de geles de electroforesis presentan carriles con varias bandas, y partiendo de la consideración

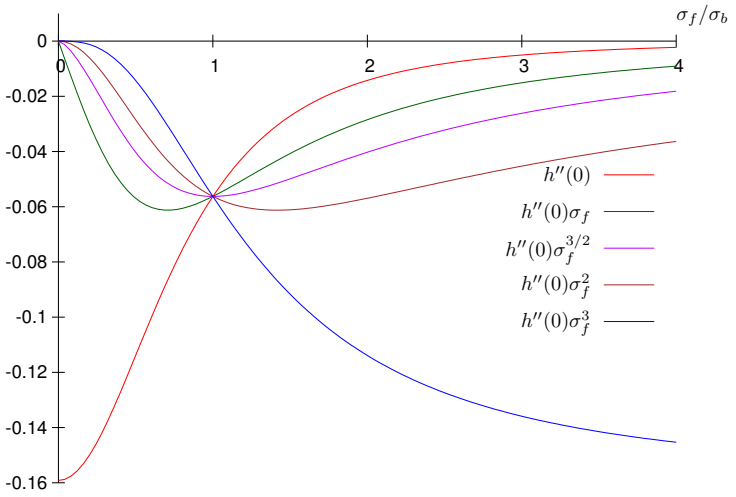


Figura 3.26: Normalización del máximo de de la segunda derivada del espacio de escalas. Normalización de (3.16) para valores de $n = 0$ (línea roja), $n = 1$ (línea verde), $n = 1, 5$ (línea magenta), $n = 2$ (línea café) y $n = 3$ (línea azul) para un $\sigma_b = 1$.

de que todas las bandas poseen la misma desviación estándar, puede encontrarse la región con la mayor cantidad de bandas de la imagen mediante una ventana.

El proceso de enventanado del espacio de escalas consiste en el cálculo del espacio de escalas tal como se muestra en la figura 3.27. El espacio de escalas es recorrido en busca del punto máximo para cada una de las bandas, identificados como puntos blancos en la imagen.

Utilizando los valores máximos se genera un vector de densidad de puntos máximos con respecto a la desviación estándar realizando una convolución con una función rectangular y luego con una función semicircular para obtener una función suavizada que facilite la búsqueda del máximo, el cual corresponde al valor de desviación estándar más probable en el espacio de escalas. En la figura 3.28 se ejemplifica este proceso de generación de una función de densidad de puntos máximos a partir de dos convoluciones, donde cada uno de los filtros aplicados tiene un ancho de $\sigma = 1$, ó 100 píxeles que es la resolución utilizada en el vector de máximos con respecto a σ obtenidos del espacio de escalas.

La aplicación de un filtro rectangular se realiza para evitar el decrecimiento de la función resultante de la convolución en la región entre máximos con poca separación al utilizar un filtro de suavizado de tipo gaussiano, semicircular ó alguna forma similar. Filtrando primero con una función rectangular es posible establecer como máximo del espacio de escalas un punto intermedio entre los dos máximos encontrados con poca separación, pero es necesario realizar un suavizado de la función que permita realizar la búsqueda del punto máximo, para lo cual en este trabajo se eligió una función semicircular. Sin embargo, para el filtrado de los datos se realiza primero la convolución entre estos dos filtros, para luego filtrar los datos en un único paso.

Una vez que se ha establecido el valor máximo se calcula una ventana centrada en ese punto, cuyo ancho es un parámetro definible por el usuario, de forma que se puedan tomar

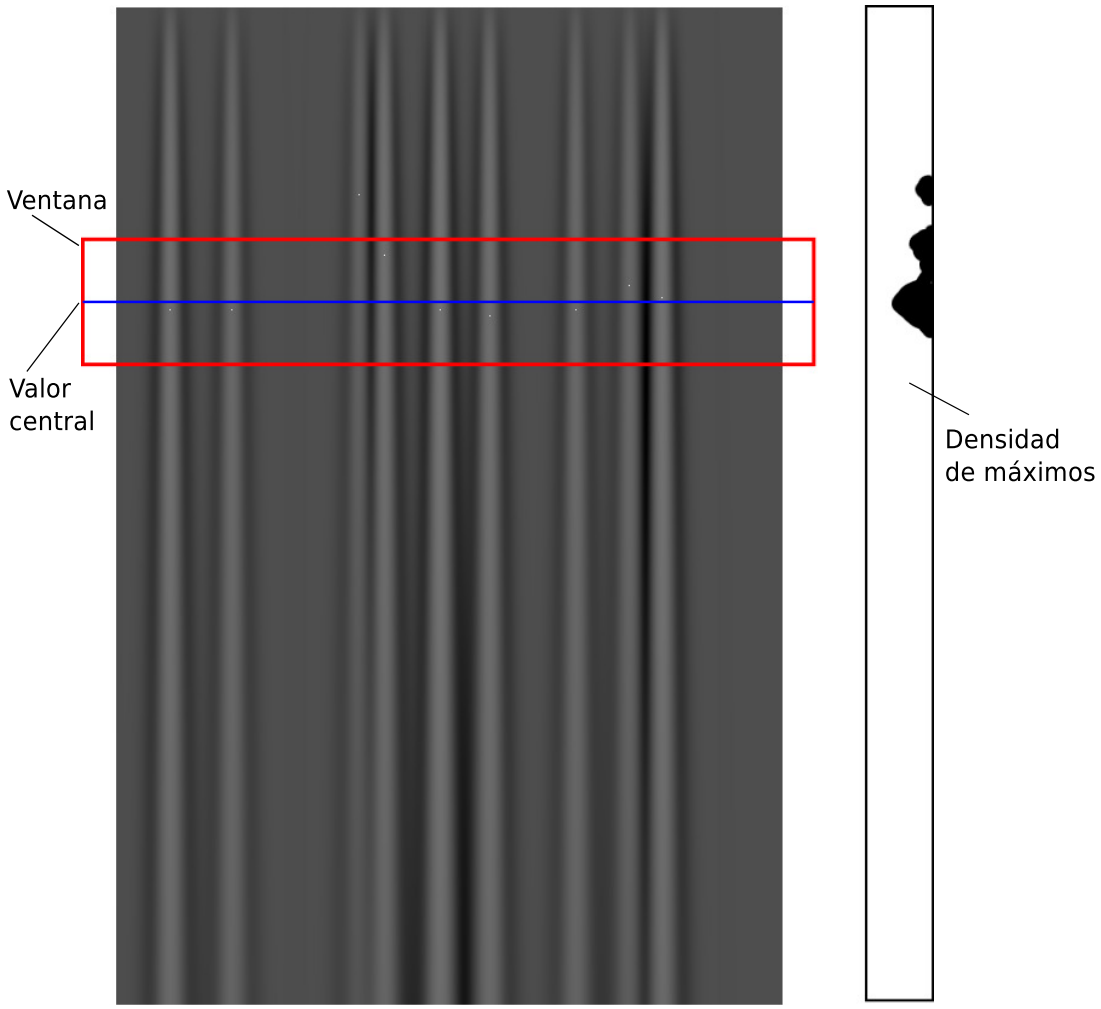


Figura 3.27: Enventanado del espacio de escalas.

en cuenta las bandas cuyo máximo se encuentra dentro de esta ventana como los datos más representativos del carril, puesto que son los puntos máximos más cercanos al valor central.

Problemas en la determinación de la desviación estándar

El espacio de escalas permite ubicar de forma correcta la ubicación de los máximos para una banda; sin embargo, puesto que la segunda derivada posee lóbulos laterales de signo opuesto se afecta la respuesta de múltiples bandas adyacentes. La influencia de este fenómeno sobre otras bandas se incrementa cuando aumenta la cantidad de bandas juntas. La influencia de las otras bandas influye tanto en la posición del máximo como en la amplitud del espacio de escalas correspondiente a cada banda.

Según [31] la desviación estándar de una sumatoria de funciones gaussianas no puede ser determinada exactamente por medio del espacio de escalas cuando existen dos ó más gaussianas traslapadas; sin embargo, puede utilizarse la desviación estándar de este método como un valor inicial para la aproximación de las funciones gaussianas mediante mínimos

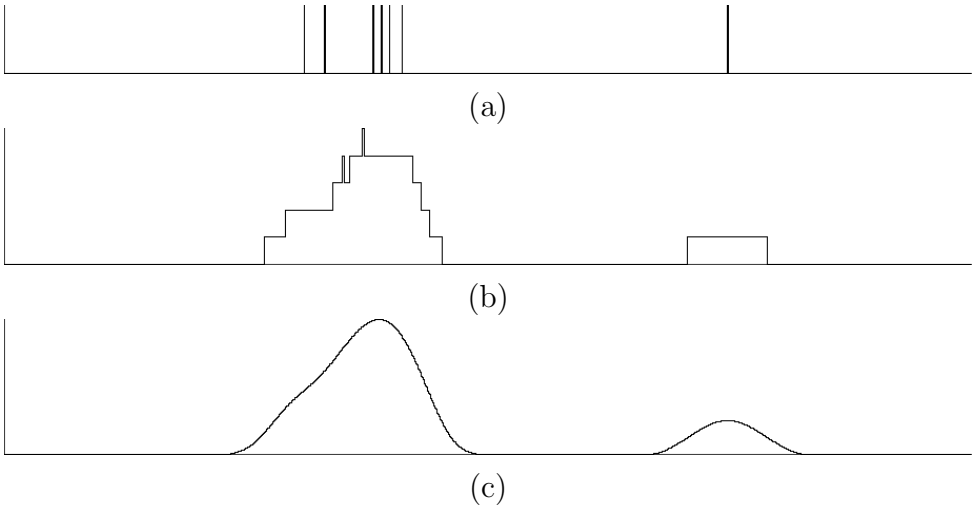


Figura 3.28: Detección de la desviación estándar utilizando espacio de escalas. (a) Distribución máximos en el espacio de escalas con respecto a σ. (b) convolución de (a) con un impulso rectangular; (c) convolución de (b) con una función semicircular.

cuadrados [53].

Mejora de la resolución utilizando derivadas

La solución planteada se basa en el modelo de bandas descritas como funciones gaussianas de igual desviación estándar, de modo que el carril se encuentra formado por la sumatoria de estas funciones. En el caso ideal en el que las funciones se encuentran separadas podría realizarse una búsqueda de máximos con el fin de encontrar la posición de las bandas una vez eliminado el ruido.

Como se muestra en la figura 3.25, en el punto máximo de una función gaussiana, la primera derivada pasa por cero cambiando su valor de positivo a negativo. La detección de este cambio es un indicio de que en este punto se encuentra un máximo.

Además la segunda derivada presenta un mínimo en la ubicación del máximo de la gaussiana de interés, por lo que el análisis de ambas derivadas es complementario en la búsqueda de los valores máximos en carriles formados por funciones gaussianas.

La detección de la ubicación de las bandas mediante el proceso de búsqueda de máximos no puede realizarse cuando se procesan imágenes en las cuales el traslape existente entre las bandas adyacentes es tan severo que es imposible distinguirlas, aún si la señal es carente de ruido, como se muestra en la figura 3.29. En esta figura se muestra un carril teórico con tres funciones gaussianas con una separación de 2 veces la desviación estándar.

La figura 3.29 muestra cómo al desaparecer los máximos locales en el carril, la primera derivada no cruza por cero en la ubicación correspondiente a los máximos de cada una de las bandas. Este hecho descarta el simple uso de la primera derivada como método para hallar los máximos.

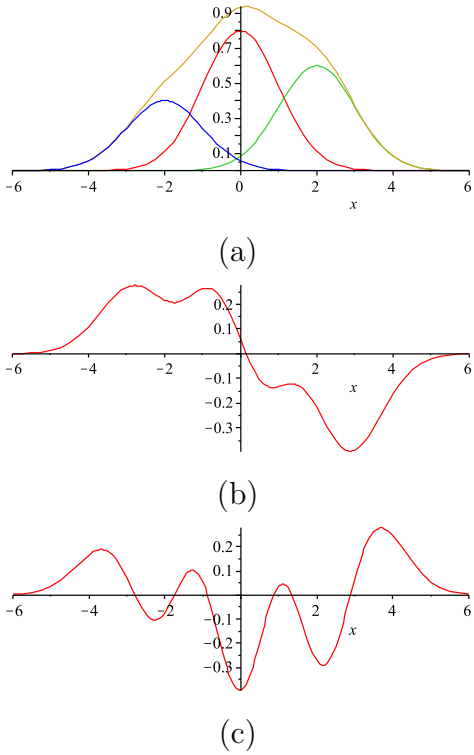


Figura 3.29: Derivadas de una sumatoria de gaussianas con traslape severo. (a) Sumatoria de gaussianas con traslape severo, (b) primera derivada, (c) segunda derivada

Sin embargo, la segunda derivada sí conserva la generación de los mínimos en una posición próxima a la de los máximos de las tres gaussianas, por lo tanto esta es más confiable para la ubicación de la posición de las bandas en un carril con bandas traslapadas.

El cálculo de la segunda derivada se realiza una única vez utilizando el filtrado y suavizado de Savitzky y Golay [53] en un único paso. Este se hace al inicio del proceso para encontrar la posición de las bandas, considerando un ancho de filtro de 3σ de acuerdo con [21], donde se recomienda utilizar para el filtrado de funciones gaussianas un ancho del filtro no mayor a 0.9 veces el ancho de las bandas a la mitad de la altura (aproximadamente $2,35\sigma$ para una distribución gaussiana), esto para producir una distorsión de la altura de los picos no mayor al 1%. Por lo tanto, para el filtrado se utiliza un ancho de 2σ. Luego de la segunda derivada se aplica un filtrado para suavizado de los datos finales utilizando el suavizado de Savitzky y Golay [53]. Tanto para la derivada como para el suavizado se utiliza un polinomio de Gram de segundo orden. Para las bandas más angostas se utiliza un ancho mínimo del filtro de 7 píxeles, puesto que para anchos menores la extracción de ruido es deficiente en imágenes reales.

En este proceso de derivación se aplica también una umbralización de los máximos encontrados en la segunda derivada, de forma que sólo serán aceptables los que superen el promedio de la parte positiva de la segunda derivada.

Modelado del carril como sistema lineal

Las bandas presentes en un carril con traslape son identificables si se considera el análisis utilizando la segunda derivada, tal como se explicó en la sección anterior. Conociendo las posiciones de cada una de las bandas se modela el carril completo ya sin fondo como una sumatoria de funciones gaussianas de diferentes amplitudes pero de igual desviación estándar, ubicadas en las posiciones μ_i encontradas con la segunda derivada. De esta forma, el carril completo $l(x)$ puede expresarse como la siguiente serie con desviación estándar constante σ

$$l(x) = \sum_{i=1}^{n} A_i \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu_i}{\sigma}\right)^2} \tag{3.17}$$

Utilizando la expresión anterior puede construirse un sistema lineal con los valores del carril sin fondo $f(x)$, evaluado en cada una de los máximos encontrados μ_i como:

$$\begin{pmatrix} f(\mu_1) \\ f(\mu_2) \\ \vdots \\ f(\mu_n) \end{pmatrix} = \begin{pmatrix} g_1(\mu_1) & g_2(\mu_1) & \cdots & g_n(\mu_1) \\ g_1(\mu_2) & g_2(\mu_2) & \cdots & g_n(\mu_2) \\ \vdots & \vdots & \ddots & \vdots \\ g_1(\mu_n) & g_2(\mu_n) & \cdots & g_n(\mu_n) \end{pmatrix} \begin{pmatrix} A_1 \\ A_2 \\ \vdots \\ A_n \end{pmatrix} \tag{3.18}$$

En este sistema lineal se expresa el valor en un punto máximo como la suma de las componentes de cada una de las otras bandas, donde $g_i(x)$ se define como

$$g_i(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu_i}{\sigma}\right)^2} \tag{3.19}$$

Una vez que se define un valor de desviación estándar σ se puede realizar el cálculo de las amplitudes A_i correspondientes a cada una de las bandas para una desviación estándar determinada resolviendo el sistema lineal anterior.

Ajuste por minimización de una función objetivo

El carril de ajuste generado en la sección anterior $l(x)$ (3.17) se aproxima al carril real sin fondo definido por una función $f(x)$ cuando la desviación estándar utilizada para la aproximación es igual al valor teórico de las bandas, en cuyo caso el error sería el mínimo. Se utiliza la función de error cuadrático medio (3.12) ya presentada en la sección 3.4.1.

3.5 Ajuste de la posición de las bandas y segmentado del carril en ventanas

Se utiliza la optimización de la posición de las bandas encontradas utilizando el algoritmo *Downhill Simplex* [53], el cual realiza una optimización multidimensional de un conjunto de los datos de posición y amplitud para un carril.

Puesto que la cantidad de iteraciones del algoritmo incrementa cuanto mayor sea la cantidad de dimensiones consideradas, se realiza la división del carril en ventanas, utilizando el método desarrollado en [59] y ya mostrado en la sección 3.4.1.

Al realizar esta división y conocerse en forma aproximada la posición de las bandas, la cantidad de iteraciones necesarias para la convergencia del algoritmo se reduce al cálculo de vectores de menos dimensiones en cada ventana. Sin embargo, este método tiene como inconveniente que las bandas encontradas en las zonas de 8σ a partir los bordes de cada ventana estarán duplicadas en la ventana adyacente, por lo que debe establecerse una forma de fusión de bandas para estos casos en los datos finales.

Una vez que las bandas detectadas en todas las ventanas son obtenidas, éstas son incluidas en un único vector y ordenadas por la posición. Si dos bandas están separadas por una distancia menor a δ las bandas pueden ser fusionadas. Para esta fusión se debe considerar que si en una de las ventanas la banda no se encuentra completa, ubicada a menos de 3σ de uno de los extremos, se descarta y se toma la banda presente en la ventana adyacente porque está aproximada considerando más información. Si la banda está aproximada con suficiente información en ambas ventanas se utiliza la siguiente fórmula para realizar la fusión de las bandas:

$$\mu_t = \frac{A_i\mu_i + A_{i+1}\mu_{i+1}}{A_i + A_{i+1}} \tag{3.20}$$

donde A_i y μ_i representan la amplitud y posición de una banda; A_{i+1} y μ_{i+1} representan la otra banda a fusionar y μ_t es la posición de la banda resultante, donde la amplitud en este punto A_t, se obtiene de leer el valor del carril en esa posición.

3.6 Extensiones de seguridad para ATEGI

3.6.1 Validación de Ingreso al Sistema

Parte importante de cualquier sistema, se basa en la seguridad que se pueda brindar, de forma que las personas puedan ingresar al sistema solamente cuando se les haya asignado un usuario y una clave, por lo que esta parte fue implementada en la etapa recién terminada del proyecto.

Se agregó una pantalla de inicio (figura 3.30) que será la encargada de hacer la validación de los usuarios y sus respectivas contraseñas, esto basado en lo implementado previamente en la base de datos.

Figura 3.30: Pantalla de inicio

En caso de fallas en el usuario o clave, el sistema le retroalimentará al usuario el respectivo error y lo redigirá nuevamente a la misma pantalla de inicio (figura 3.31).

Figura 3.31: Error en identificación

Se agregaron funcionalidades a la pantalla de configuraciones para poder tener acá acceso a los mantenimientos de Compañías, Usuarios, Marcas, Modelos, Equipo (figura 3.32).

Figura 3.32: Accesos a mantenimiento de datos

Para las compañías se agregó el mantenimiento en la figura 3.33. Las compañías tienen un identificador único que es un auto incremental. De este lado se puede consultar las

Figura 3.33: Mantenimiento de compañías

idCompania	Nombre	Contacto	Telefono	fax	Telefono Contacto	Direccion
28	a	a	a	a	a	a
29	b	b	b	b	b	b
30	ege	k	j	n	d	j
31	ege	k	j	s	d	j
32	4	4	4	4	4	4
33	9	9	9	9	9	9
34	Andres Co	Andres Gonzales	88888888	22222222	77777777	[illegible]
35	2	2	2	2	2	2
36	Andres Co					
37	Compañía01	Representante 01	Teléfono 01	Fax 01	Teléfono 01	Dirección 01
38	Compañía01	Representante 01	Teléfono 01	Fax 01	Teléfono-Representante 0	Dirección 01
39	Compañía01	Representante 01	Teléfono 01	Fax 01	Teléfono-Representante 0	Dirección 01
40	Compañía01	Representante 01	Teléfono 01	Fax 01	as	Dirección 01
41	Compañía01	Representante 01	Teléfono 01	Fax 01	as	Dirección 01

Figura 3.34: Listado de compañías

compañías existentes en la base de datos (figura 3.34).

En el área de mantenimiento de usuarios (figura 3.35) estos se registran de acuerdo a las compañías anteriormente ingresadas.

Figura 3.35: Mantenimiento de usuarios

Se pueden ver los usuarios creados, cada usuario tiene un identificador único (figura 3.36).

Figura 3.36: Listado de usuarios

Las marcas y modelos de dispositivos también deben tener su mantenimiento, por lo que se crearon acá de forma que cuando en el análisis se realice, se pueda registrar el equipo, marca y modelo en el que se realizó la corrida (figuras 3.37-3.40).

Una vez definidas las marcas y modelos, se puede proceder a definer los equipos, de forma que todos los datos queden completamente correlacionados (figura 3.41).

3.7 Extensión de la arquitectura del sistema para manejar bases de datos distribuidas

El objetivo de la propuesta original del proyecto perseguía la implementación de una extensión de la arquitectura de bases de datos del sistema para permitir combinar la

Figura 3.37: Mantenimiento de marcas y modelos de dispositivos

Figura 3.38: Lista de marcas

Figura 3.39: Agregar nuevo modelo

Figura 3.40: Lista de modelos

Figura 3.41: Creación de equipos

información de múltiples sitios con bases de datos de geles. Debido a recortes presupuestarios en la aprobación del proyecto, este objetivo se redujo a solo el diseño de dicho sistema de bases de datos distribuidas.

En el estado del arte (sección 2.7) se presentó un análisis de arquitecturas distribuidas para determinar el mejor enfoque dado las características del sistema en desarrollo. A continuación se presenta el diseño de un esquema de exportación que modela la información que será compartida entre los participantes del sistema distribuido. Después se describe cómo se procesarían las consultas y se daría mantenimiento a un mecanismo de caché para reducir el intercambio de datos. Se finaliza presentando el protocolo que se seguirá para que nuevos nodos se incorporen al sistema distribuido.

3.7.1 Requerimientos generales

- Permitir realizar búsqueda de geles y carriles en las bases de datos de los nodos participantes. La información extraída en las búsquedas debe permitir realizar tareas de minería de datos sobre los carriles.

- Se debe mantener independencia de los participantes en cuanto al manejo de su información de geles, tanto para el contenido que comparte como en el manejo de información adicional no compartida.

- Reducir el tráfico de red.

3.7.2 Estructuras de almacenamiento de información

Cada nodo ofrecerá la información almacenada en un cubo de datos diseñado para responder eficientemente a consultas sobre: geles, carriles, muestras y equipos usados. Cada tuple del cubo incluirá adicionalmente un campo con la fecha del dato. Dicha fecha se define como la fecha más reciente entre las fechas de los datos usados para obtener ese tuple. El modelo de la base de datos requerirá incluir dicha fecha como un cambio en el diseño de las siguientes tablas: `tiposGel`, `gel`, `carrilesGel`, `bandas`, `equipo`, `tipoEquipo`, `modelos`, `marcas`, `muestras` y `origen`.

El proceso de generación del cubo se encargará de asignarle a cada tuple del cubo la fecha de modificación más reciente de entre las fechas de modificación de los tuples usados para generar ese tuple del cubo.

Adicionalmente, cada nodo contendrá un caché con los tuples que ha obtenido de otros nodos como respuesta a consultas previas. Además de la información base del cubo, el caché contendrá un identificador del nodo que contiene la información original, la fecha en modificación del dato y la fecha de expiración del registro del caché. El sistema eliminará periódicamente del caché aquellos registros cuya fecha de expiración se haya cumplido. La fecha de expiración se calculará en base a un parámetro con los días de validez. Cada

vez que un tuple es obtenido de su nodo o cada vez que se verifica su validez, se recalcula la fecha de expiración.

Asimismo cada nodo mantiene un registro de los tuples de su cubo que ha enviado a otros nodos de la red. Para ello mantiene una tabla en que se registra el identificador del nodo al que se le envió la información, el identificador del tuple enviado (`idgel`, `idcarril`) y la fecha del dato. Esta tabla será usada para evitar enviar información a un nodo que ya la debería tener.

3.7.3 Procesamiento de consultas

Los nodos envían consultas a otros nodos para extraer información de los cubos de esos otros nodos. Para cada tuple de la respuesta se revisa si el nodo solicitante tiene una copia de ese tuple que todavía es válida. En estos casos, dichos tuples no son enviados completos sino que se incluyen sus identificadores una lista que se adjuntará a la respuesta.

MySql permite ejecutar una instrucción que produce múltiples conjuntos de resultados [50]. De modo que como resultado de una consulta, los nodos pueden devolver los tuples completos para aquellos tuples no conocidos previamente por el nodo que consulta, seguido de una lista de identificadores de tuples ya conocidos por el nodo que consulta. La figura 3.42 ilustra el proceso de consulta de información.

3.7.4 Mantenimiento del cubo

Al calcular o actualizar el cubo se debe determinar qué tuples han variado con el fin de determinar al responder consultas cuáles tuples de las respuestas ya están en poder del nodo que consulta. Para lograr lo anterior, se tendrá una columna llamada `FechaDato` en los tuples del cubo, cuyo valor será el valor máximo de las columnas `FechaModif` de las tablas usadas para generar el cubo. De esa manera, si ninguno de los datos originales usados en la construcción de un tuple del cubo ha sido modificado, su `FechaDato` se mantendrá igual; por el contrario, cualquier actualización de los datos originales, posterior a la generación anterior del cubo, producirá una `FechaDato` distinta.

3.7.5 Mecanismo de inscripción

MySql establece la restricción de que para acceder a una base de datos desde un servidor diferente al usado para correr el motor de la base de datos, se debe tener un usuario con password junto con la dirección IP del servidor externo. Esto vuelve mucho más rígidas las comunicaciones entre nodos porque al no haber un mecanismo de acceso público general, los nodos si quieren compartir información deben conocerse mutuamente.

De modo que para que un nodo pueda hacer una consulta que corra en todos los nodos de la red, necesita primero contactarlos para que cada uno de dichos nodos le dé un usuario

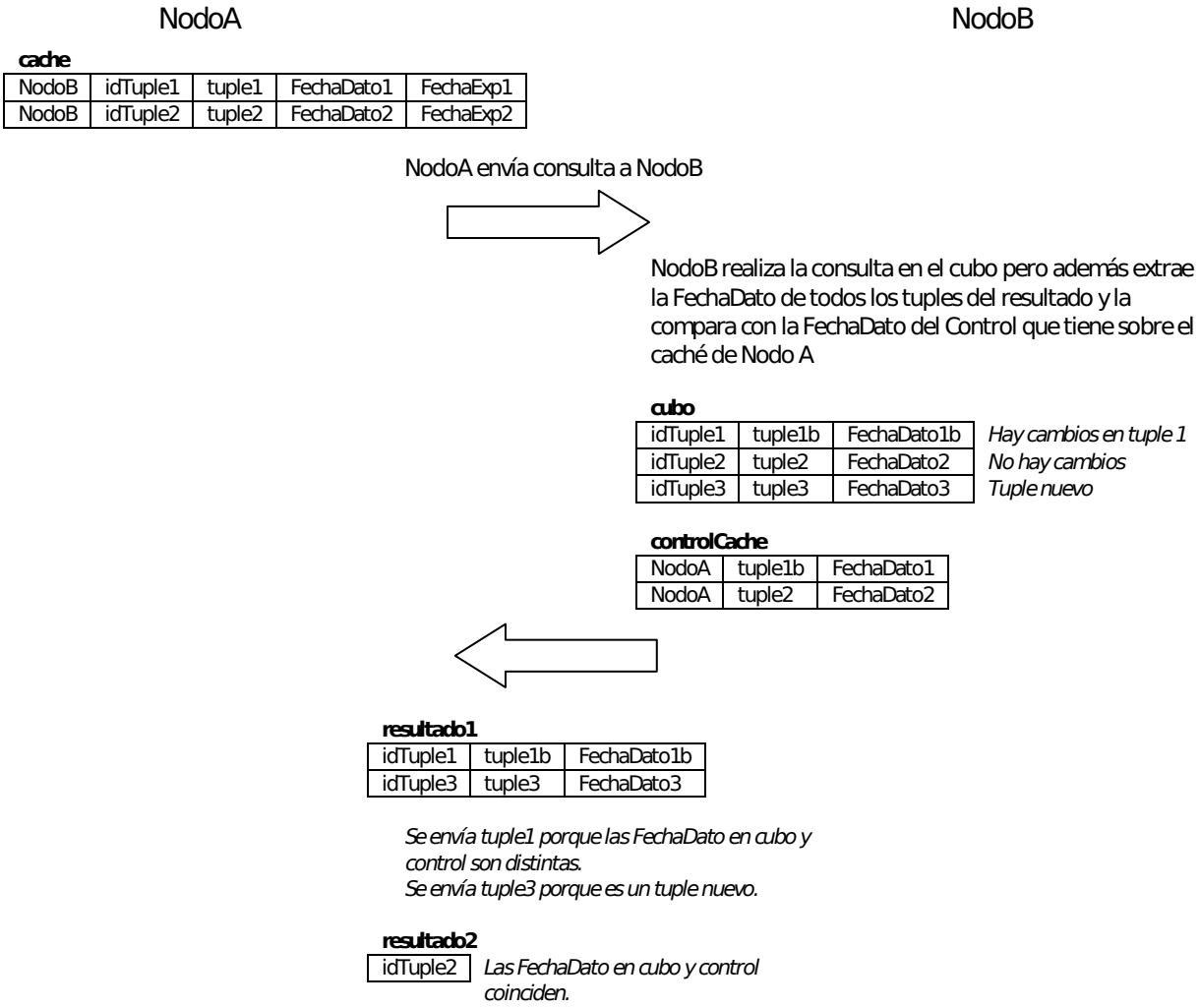


Figura 3.42: Proceso de consulta de información.

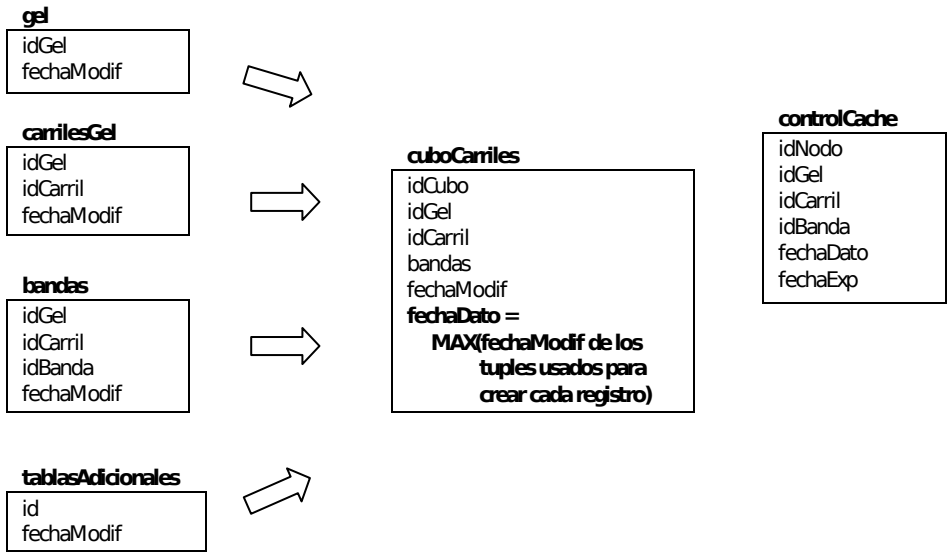


Figura 3.43: Mantenimiento del cubo.

y password.

Para estructurar el mecanismo de inscripción, se tendrán nodos especiales, llamados súper pares, que servirán de enlace inicial a los nuevos nodos que quieran incorporarse. Los súper pares forman una subred en la que todos se conocen y mantienen replicada y sincronizada la información sobre todos los demás nodos de la red. Los súper nodos le dan a los nuevos nodos una lista con la dirección URL de todos los demás participantes del sistema distribuido. Usando esas direcciones los nuevos nodos contactan a esos participantes y obtienen las credenciales (usuario/password) que les permita acceder a la información provista.

El mecanismo de incorporación sería el siguiente:

1. En el instalador de la herramienta viene incluida una lista con las direcciones IP de los súper pares conocidos al momento de crear el instalador.

2. El nodo nuevo escoge aleatoriamente un súper par y lo contacta para solicitar la incorporación a la red. Si el contacto falla y el súper par no responde que recibió la solicitud, el nodo nuevo intenta con otro súper par escogido aleatoriamente.

3. El súper par contactado confirma inicialmente la recepción de la solicitud. Una vez que el administrador de súper par autoriza la incorporación, el súper par le informa al nuevo nodo del resultado y en caso de ser aceptado le solicita un usuario/password para verificar acceso.

4. El nuevo nodo al recibir la autorización del súper par, crea una cuenta con usuario/password para dicho nodo en MySql y le notifica.

5. El súper par contactado recibe el usuario/password enviado por el nuevo nodo y verifica que efectivamente se tiene acceso a los datos de ese nuevo usuario. Una

vez verificado el acceso, el súper par crea una cuenta para el nuevo nodo (usuario, password) y registra su IP con MySql. Luego de creada la cuenta el súper par le envía al usuario los datos de su cuenta, y le envía además una lista con las direcciones de los demás participantes, tanto otros súper pares como nodos comunes.

6. El nuevo nodo recibe la respuesta del súper par. Verifica el acceso provisto por el súper par con usuario/password. Luego procede a tomar la lista de los demás participantes y los contacta para intercambiar cuentas de MySql. Cada nodo individualmente está en el derecho a negarse a autorizar el acceso al nuevo nodo. En ese caso el nuevo nodo simplemente no registra el usuario/password de los nodos que se negaron a darle acceso.

Un nodo simple puede convertirse en súper par mediante una solicitud a los otros súper pares. Se utiliza algún mecanismo de elección para que los súper pares decidan entre todos aceptar o rechazar la solicitud de un nuevo súper par. En caso de ser aceptado se sincroniza la información con el nuevo súper par.

Si bien el diseño anterior tiene problemas de escalabilidad ya que todos deben comunicarse con todos, se puede usar en el caso de un número razonable de nodos puesto que es relativamente sencillo además que no vuelve excesivamente complejas las labores de los súper pares.

3.7.6 Mecanismo alternativo de inscripción

Un diseño más escalable requiere aumentar la complejidad de los súper pares para que los mismos se encarguen de direccionar consultas a subconjuntos disjuntos de los nodos participantes. De esa manera los contactos se reducen a nodo – súper par y súper par – súper par y no la totalidad de nodo – nodo. Si hay n nodos de los cuáles k son súper pares, el primer diseño requiere $\mathcal{O}(n^2)$ contactos, mientras que el segundo, suponiendo que cada nodo tiene contacto con todos los súper nodos, requiere $\mathcal{O}(kn + k^2)$.

Como en el caso anterior, Se tendrán nodos especiales llamados súper pares que servirán de enlace inicial a los nuevos nodos que quieran incorporarse. Los súper pares forman una subred en la que todos se conocen y mantienen replicada y sincronizada la información sobre todos los nodos de la red. Adicionalmente, cada nodo ha intercambiado cuentas de acceso con dos o tres súper pares.

En este caso el mecanismo de incorporación sería el siguiente:

1. En el instalador de la herramienta viene incluida una lista con las direcciones IP de los súper pares conocidos al momento de crear el instalador.

2. El nodo nuevo escoge aleatoriamente un súper par y lo contacta para solicitar la incorporación a la red. Si el contacto falla y el súper par no responde que recibió la solicitud, el nodo nuevo intenta con otro súper par escogido aleatoriamente.

3. El súper par contactado confirma inicialmente la recepción de la solicitud. Una vez que el administrador de súper par autoriza la incorporación, el súper par le informa al nuevo nodo del resultado y en caso de ser aceptado le solicita un usuario/password para verificar acceso.

4. El nuevo nodo al recibir la autorización del súper par, crea una cuenta con usuario/password para dicho nodo en MySql y le notifica.

5. El súper par contactado recibe el usuario/password enviado por el nuevo nodo y verifica que efectivamente se tiene acceso a los datos de ese nuevo usuario. Una vez verificado el acceso, el súper par crea una cuenta para el nuevo nodo (usuario, password) y registra su IP con MySql. Luego de creada la cuenta el súper par le envía al usuario los datos de su cuenta, y le envía además una lista con las direcciones de los demás participantes, tanto otros súper pares como nodos comunes.

6. El nuevo nodo recibe la respuesta del súper par. Verifica el acceso provisto por el súper par con usuario/password. Luego procede a tomar la lista de los demás participantes y los contacta para intercambiar cuentas de MySql. Cada nodo individualmente está en el derecho a negarse a autorizar el acceso al nuevo nodo. En ese caso el nuevo nodo simplemente no registra el usuario/password de los nodos que se negaron a darle acceso.

3.8 Módulo adicional: Minería de datos

El tercer objetivo persigue agregar a la interfaz de usuario modos adicionales para la manipulación de las imágenes, la interacción con el sistema y la presentación de la información almacenada y generada.

3.8.1 Datos de entrada (cubo OLAP)

Con el fin de simplificar las búsquedas y además estructurar el módulo de modo que facilite el acceso distribuido en un futuro, el módulo de minería de datos trabaja sobre un cubo OLAP obtenido a partir de la información almacenada en la base de datos de geles.

Aunque no fue incluida en esta implementación, el proceso de generación del cubo puede fácilmente incluir información de control en cuanto al nivel de seguridad que debe tener un usuario para poder acceder a la información del cubo.

El cubo se crea combinando información de las siguientes tablas: tiposgel, gel, carrilesgel, bandas, muestras, origen, equipo, tipoequipo, modelos, marcas. El contenido de la tabla bandas fue desnormalizado: para cada carril se obtuvieron todas sus bandas y se agruparon en un campo de texto que consiste en una secuencia de números de bases separados por caracteres '|'. Esta de normalización facilita obtener la información de los carriles que es necesaria para alimentar a los algoritmos de minería de datos.

El cubo permite seleccionar carriles con base a la información almacenada sobre los geles, equipos, muestras para así obtener las bandas correspondientes y aplicar algoritmos de minería de datos usando esas bandas como características. El diseño del cubo sigue un modelo en estrella tal como se muestra en la figura 3.44.

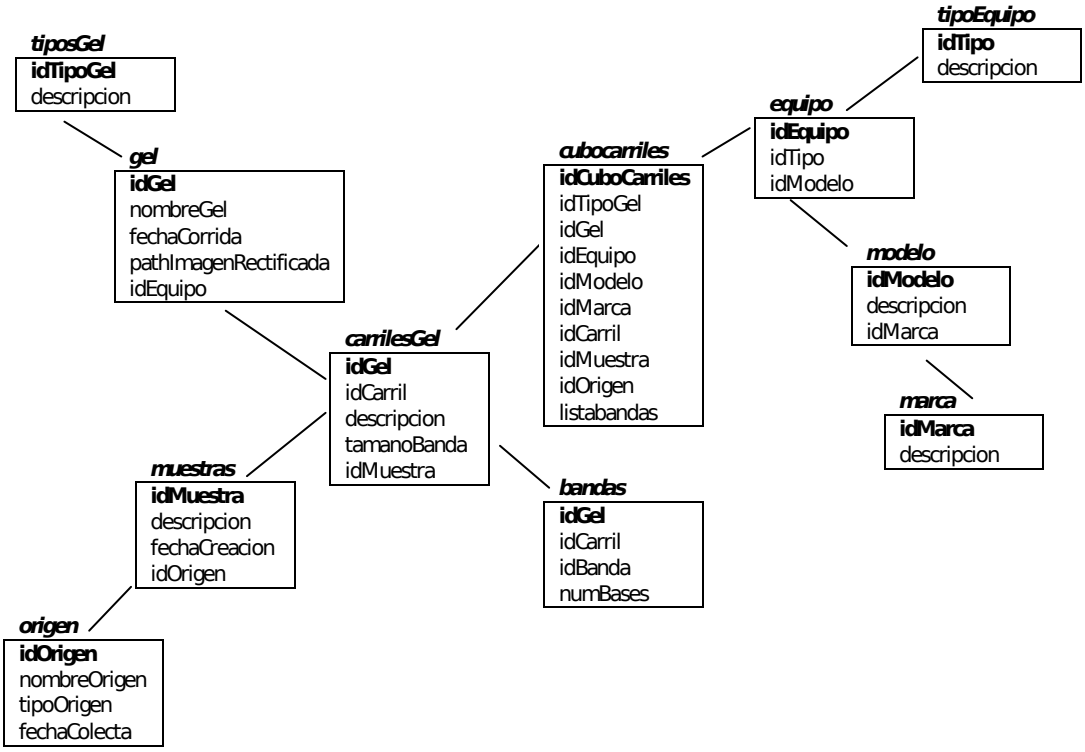


Figura 3.44: Diseño del cubo con modelo en estrella.

3.8.2 Proceso de creación del cubo

El cubo es creado mediante la ejecución de un procedimiento almacenado dentro de la base de datos llamado `CreaCuboCarriles`. Para todos los carriles que hay en la base de datos, este procedimiento combina la información respectiva de geles, tipoequipo, equipo, modelos, marcas, muestras, origen. Además, combina la información de las bandas de un carril en un campo de texto en que los pesos moleculares se presentan en orden ascendente y separados por un carácter especial. La figura 3.45 ilustra el proceso.

3.8.3 Funcionalidad implementada: clasificación y clustering

El módulo de minería de datos provee dos técnicas muy populares: clasificación por árboles de decisión y agrupamiento (clustering). La herramienta Weka implementa varios algoritmos de clasificación, de los cuales el módulo de minería de datos permite usar el método J48. Dicho método es una versión posterior del muy popular algoritmo C4.5 desarrollado por J. Ross Quinlan [55].

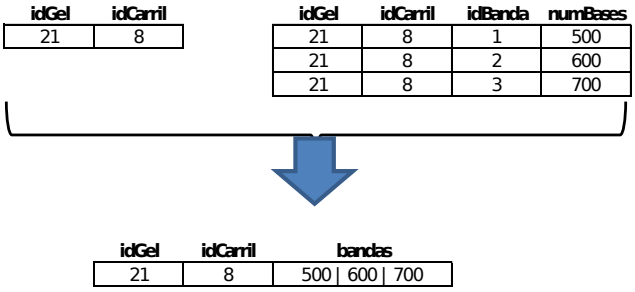


Figura 3.45: Proceso de creación del cubo.

Weka implementa varios algoritmos de clustering como k-means, EM, o Cobweb. De estos el módulo de minería de datos provee el método k-means, pero es fácilmente adaptable para que pida parámetros e invoque otros métodos de clustering.

3.8.4 Interfaz del módulo de Minería de Datos

Lo primero que se hace es escoger el tipo de gel. Para eso se usa el primer elemento web que es un campo de texto (figura 3.46).

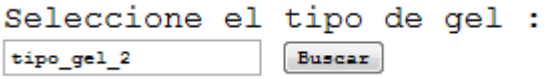


Figura 3.46: Selección de tipo de gel

Luego de encontrar los geles con el tipo adecuado se escoge uno de ellos usando el elemento de la página web ilustrado en la figura 3.47.

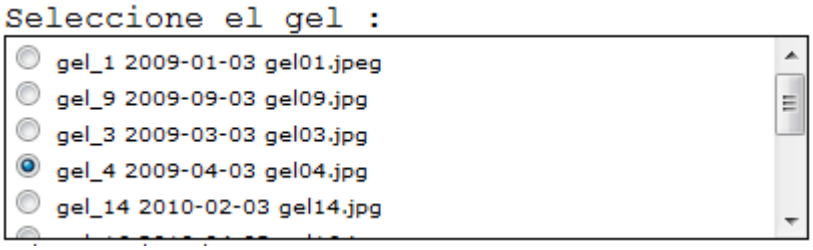


Figura 3.47: Selección de tipo de gel

Al escoger uno de los geles la interfaz muestra una imagen del mismo (figura 3.48).

También se muestra una lista de los carriles del gel escogido en un formulario que permite escoger los carriles que serán incluidos en el análisis (figura 3.49). Si el análisis es una clasificación, además de escoger los carriles, se les debe asignar una clase a cada carril.

Los carriles escogidos no serán tomados en cuenta por las herramientas de análisis a menos que su selección sea guardada. Para lograr esto, se usa el botón `Datos Seleccionados` (figura 3.50). Seleccionar un carril significa que el sistema recuerda que el carril fue

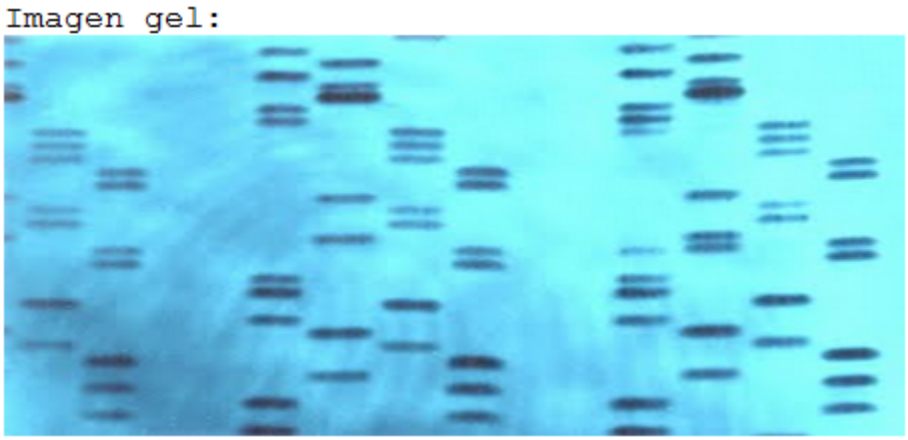


Figura 3.48: Interfaz mostrando imagen de gel seleccionado.

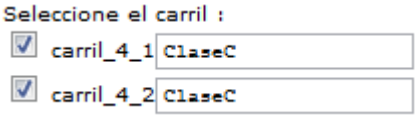


Figura 3.49: Selección de carriles.

escogido para uso de las herramientas de minería de datos. Si se quieren agregar más carriles para que las herramientas los consideren, simplemente se escogen. También es posible eliminar un carril previamente seleccionado; para eso simplemente se obtiene de nuevo y se borra el marcador de selección.

Con el fin de asegurarse de que pequeñas diferencias en los valores de una misma banda en dos o más carriles impidan a los algoritmos de análisis usar adecuadamente esa banda, el usuario tiene la posibilidad de revisar en paralelo los valores de las bandas de los carriles escogidos y combinar aquellos casos que considere sean los mismos.

Para hacer lo anterior, se presenta una lista con todos los pesos de las bandas en orden ascendente y con un campo de check, el cual si es escogido indica que ese valor se considera igual al valor siguiente.

El módulo de minería de datos antes de hacer el análisis combina los valores que han sido marcados como iguales y presenta como único representante ante los algoritmos de minería de datos al valor más pequeño. Por ejemplo, si se tienen pesos 4010, 4309, 4608,

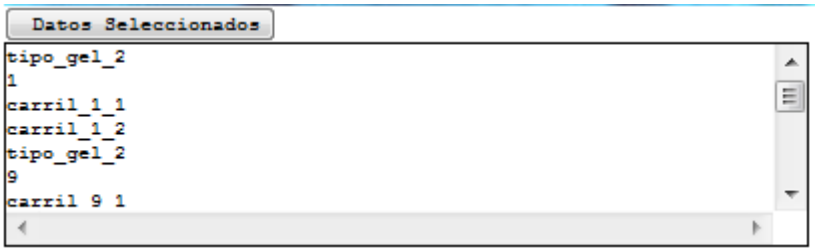


Figura 3.50: Lista de datos seleccionados.

4907 y el usuario marca los tres primeros, entonces el sistema combina esos valores y si alguno de ellos aparece en un carril el sistema lo interpreta como si apareciera la banda con peso 4010 (figura 3.51).

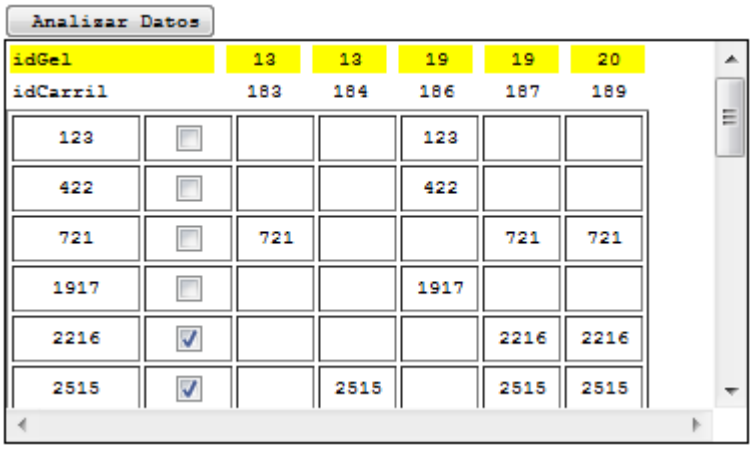


Figura 3.51: Combinación de valores marcados.

Luego de haber seleccionado los carriles de interés y de haber estandarizado los valores de los pesos de las bandas, se procede a escoger la operación de minería de datos deseada, así como a especificar sus parámetros. Las opciones disponibles son `Clustering` para hacer agrupamiento de carriles, y `Clasificación` para elaborar un árbol de decisión que clasifique los carriles.

A continuación se describen con más detalle ambas operaciones.

Clustering

Si la operación a realizar es un agrupamiento de carriles, el módulo muestra un formulario con dos parámetros (figura 3.52):

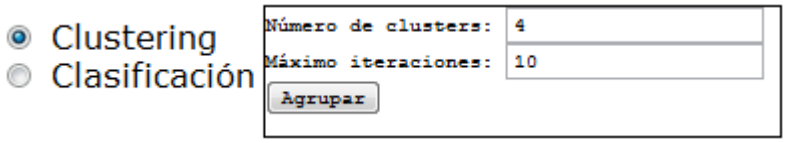


Figura 3.52: Parámetros del proceso de clustering.

- Número de clusters que se piden al algoritmo k-means.
- Número máximo de iteraciones

La salida del proceso de clustering incluye una lista de los atributos que fueron usados en el proceso, así como las instancias agrupadas (figura 3.53). Luego muestra los clusters asociados a cada instancia y finalmente los centroides de cada cluster.

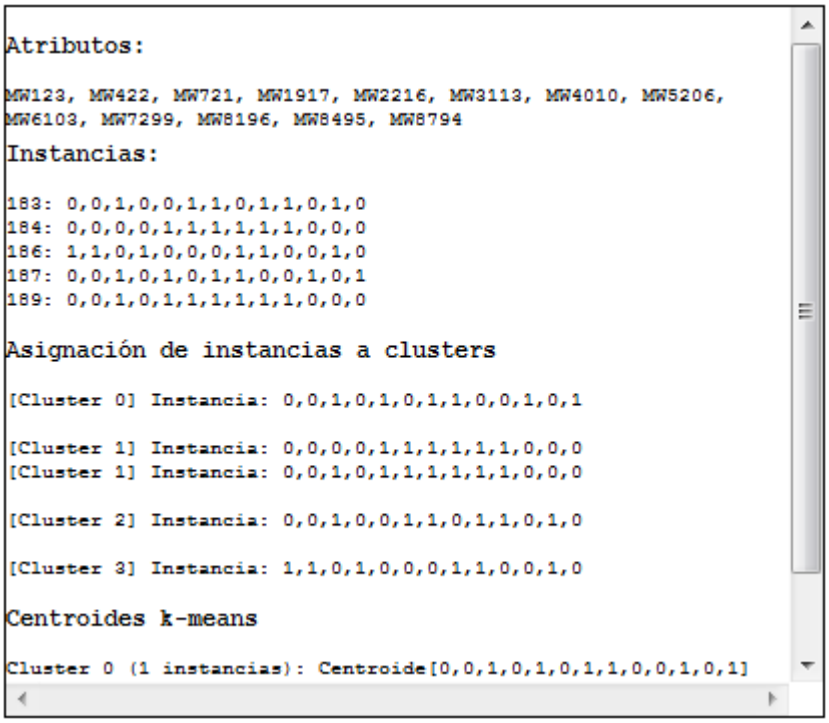


Figura 3.53: Salidas del proceso de clustering.

Clasificador J48

Si la operación a realizar es una clasificación carriles, el módulo permite ajustar varios parámetros. La escogencia de los valores de estos parámetros es muy importante porque tienen mucho impacto en la calidad de los resultados obtenidos.

Los parámetros disponibles son los siguientes (figura 3.54):

Binary splits: indica si se va a generar un árbol binario o no.

Min Num Obj: indica el número mínimo de objetos que deben haber en cada rama de una bifurcación.

Num Folds: número de fragmentos (folds) usados en la validación cruzada del modelo generado.

Reduced Error Pruning: aplicar o no una forma simple de poda en la que cada nodo, empezando por las hojas, es remplazado por su clase más popular; si no se deteriora la capacidad de predicción el cambio es aceptado

Subtree Raising: indica si se usa una forma de poda en la que un nodo se mueve hacia arriba, hacia la raíz del árbol, reemplazando a los nodos de su ruta

Unpruned: indica si se aplica poda o no al árbol; la poda debe reducir el tamaño del árbol sin que se deteriore su capacidad de predicción

Use Laplace: indica si se usa Laplace "smoothing" para predecir probabilidades evitando que sean cero.

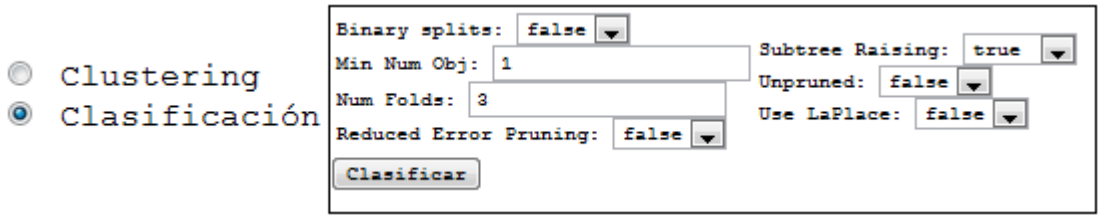


Figura 3.54: Parámetros diponibles para proceso de clasificación.

La salida del proceso de clasificación J48 incluye una lista de todas las diferentes categorías, una lista de los atributos que fueron usados en el proceso, así como una lista de las instancias clasificadas (figura 3.55). Todo esto seguido por el árbol de decisión generado, seguido de sus estadísticas de evaluación. El sistema muestra estadísticas de evaluación

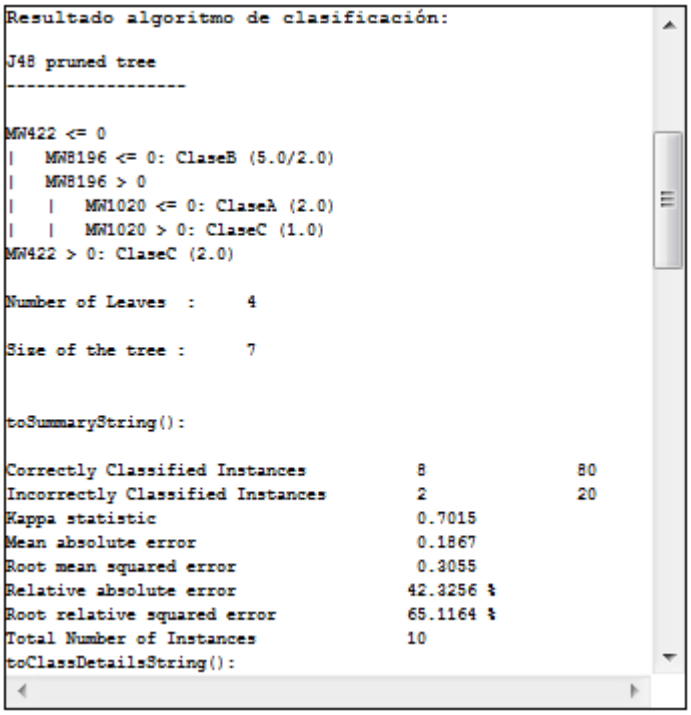


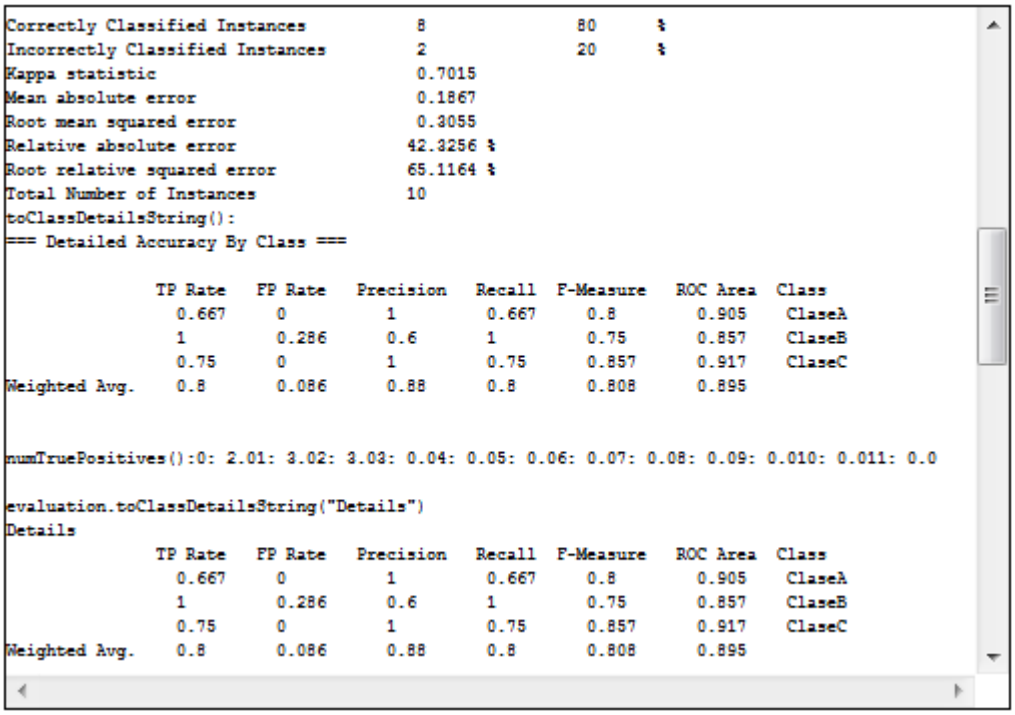
Figura 3.55: Salidas del proceso de clasificación.

totales y por clase (figura 3.56).

Al final se presenta el código Java de una clase que implementa el árbol de decisión.

3.8.5 Implementación de módulos

La implementación combinó varias herramientas de desarrollo web: Php, Javascript, Java servlets. La estructura general de la interfaz y la interconexión con la base de datos fue



```
Correctly Classified Instances           8               80      %
Incorrectly Classified Instances         2               20      %
Kappa statistic                          0.7015
Mean absolute error                      0.1867
Root mean squared error                  0.3055
Relative absolute error                 42.3256 %
Root relative squared error             65.1164 %
Total Number of Instances               10
toClassDetailsString():
=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.667    0          1         0.667     0.8         0.905    ClaseA
                 1        0.286      0.6       1         0.75        0.857    ClaseB
                 0.75     0          1         0.75      0.857       0.917    ClaseC
Weighted Avg.    0.8      0.086      0.88      0.8       0.808       0.895


numTruePositives():0: 2.01: 3.02: 3.03: 0.04: 0.05: 0.06: 0.07: 0.08: 0.09: 0.010: 0.011: 0.0

evaluation.toClassDetailsString("Details")
Details
               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.667    0          1         0.667     0.8         0.905    ClaseA
                 1        0.286      0.6       1         0.75        0.857    ClaseB
                 0.75     0          1         0.75      0.857       0.917    ClaseC
Weighted Avg.    0.8      0.086      0.88      0.8       0.808       0.895
```

Figura 3.56: Estadísticas de evaluación totales y por clase.

desarrollada en php. Por otro lado se usó Javascript para generar los elementos de la interfaz que dependen de los datos obtenidos de la base de datos así como para invocar a los servlets de Java. Finalmente se programaron en Java clases de servlets para extraer los carriles de la base de datos de acuerdo con los parámetros escogidos, preparar la información e invocar a las clases de Weka para hacer los análisis correspondientes. Los servlets de Java son provistos por el servidor web Tomcat y la programación de las clases fue hecha usando NetBeans 7.

3.9 Implementación de la DGGE para el análisis de la diversidad genética bacteriana en muestras ambientales

3.9.1 Colecta de muestras ambientales de suelo y agua

Con la finalidad de valorar los procedimientos para la toma de muestras y extracción de ADN de material ambiental se realizaron colectas tanto de suelo como de agua de distintas localidades. Las muestras de suelo fueron tomadas entre 5 y 20 cm de profundidad utilizando instrumentos estériles y luego trasladadas al laboratorio para ser tratadas y/o almacenadas a -20°C. Para el muestreo ambiental de agua se tomaron muestras a una profundidad mínima de 20 cm a partir de la superficie, abriendo y cerrando la botella estéril por debajo del agua. Las muestras fueron trasladadas inmediatamente al labora-

torio donde fueron filtradas a través de filtros con poros de 0,22 µm de diámetro con la finalidad de capturar las células procariotas. Esta muestra representaría la comunidad total microbiana *in situ*. Los filtros se mantuvieron congelados a -70°C hasta la extracción de ADN.

3.9.2 Extracción de ADN de suelos

El ADN total de las muestras de suelo fue extraído a partir de aproximadamente 250 mg de suelo utilizando el kit comercial *PowerSoil DNA Isolation Kit* (MO BIO Laboratories, Inc.) siguiendo las instrucciones del fabricante. El kit posee la capacidad de recuperar ADN aún en ambientes con altos contenidos de ácidos húmicos, eliminando a la vez posibles sustancias inhibidoras de PCR [24].

3.9.3 Extracción de ADN de muestras de agua

En el caso de la extracción de ADN de las muestras ambientales de agua, se tomó el filtro de cada muestra y se cortó en trozos de aproximadamente 0,2 cm^2. Este material se utilizó para extraer el ADN utilizando igualmente el kit comercial *PowerSoil DNA Isolation Kit*.

3.9.4 Análisis de la integridad y la cantidad de ADN genómico extraído

Se prepararon geles de agarosa al 1% en búfer TAE 1X conteniendo 0,1% de bromuro de etidio. De cada ADN extraído se cargaron en los geles 8 µL más 2 µL de búfer carga 6X. Se emplearon 5 µL de marcador de peso molecular *MassRuler DNA LadderMix* de Fermentas®. El material fue evaluado por electroforesis a 100 V por 45 minutos.

3.9.5 Reacción en Cadena de la Polimerasa (PCR)

Al ADN extraído de las muestras ambientales se le realizaron PCR utilizando imprimadores DGGE universales para el gen 16S del ARNr de eubacterias. La pareja de iniciadores utilizados fue el 341F (CCTACGGGAGGCAGCAG) con una *5'GC clamp*: CGCCCGCCGCGCGCGGCGGGCGGGGCGGGGGCACGGGGGG y el 534R (ATTACCGCGGCTGCTGG). Las reacciones de PCR fueron realizadas en un volumen total de 50 µL de acuerdo al siguiente protocolo: BufferDreamTaq™1X de Fermentas®, 0,5 µM de cada iniciador, 200 µM de una mezcla de desoxinucleósidostrifosfatos, 1,5 U de DreamTaqTMADN polimerasa (Fermentas®) y 5 µL del ADN extraído. El perfil térmico utilizado para la amplificación fue: una desnaturalización inicial (94°C, 5 min), seguido de 20 ciclos de 94°C, 45 s; 65°C, 45 s; y 72°C, 2 min con un decrecimiento de la temperatura de hibridación de 0,5°C por ciclo. Esto fue seguido por 20 ciclos de 94°C, 30 s; 55°C, 30 s; y

72°C, 2 min; más un paso de extensión final de 10 min a 72°C [49]. La amplificación de los productos de PCR del tamaño correcto fue confirmado por electroforesis en geles de agarosa al 1,5% en búfer TAE 1X conteniendo 0,1% de bromuro de etidio. El tamaño de los fragmentos de ADN esperados son de aproximadamente 233 pb [48]. La corrida electroforética fue realizada a 80 V por 45 min.

3.9.6 Electroforesis en Gel de Gradiente Desnaturalizante (DGGE)

La DGGE fue realizada utilizando el sistema *DCode Universal Mutation Detection System* (Bio-Rad). Los productos de la PCR (10 µL) fueron directamente aplicados en geles de poliacrilamida al 8% (p/vol) (acrilamida-N,N'-metilenbisacrilamida, 37, 5:1) con un tamaño de 16×16 cm, 1 mm de grosor y con un gradiente lineal de desnaturalización del 40% al 65%. Los geles de tipo paralelo fueron elaborados utilizando un aparato de formador de gradiente (*Model 475 GradientDeliverySystem*, Bio-Rad). Para esto se prepararon soluciones madre de acrilamida al 8% con 40% y 65% de agentes desnaturalizantes. La solución madre al 40% desnaturalizante fue preparada agregando 20 ml de una solución de acrilamida al 40%, 2 ml de búfer TAE 50X, 16 ml de formamida (desionizada), 16,8 g de urea y aforada con agua destilada a 100 ml. Para preparar la solución madre al 65% desnaturalizante se adicionó 20 ml de una solución de acrilamida al 40%, 2 ml de búfer TAE 50X, 26 ml de formamida (desionizada), 27,3 g de urea y finalmente aforada con agua destilada a 100 ml. Ambas soluciones madre fueron desgaseadas por 10 a 15 minutos utilizando una bomba de vacío, filtradas a través de filtros de 0,45 µm y mantenidas a 4°C en botella ámbar para su uso. Las electroforesis fueron realizadas en búfer TAE 1X (40 mM Tris pH 8,1; 20 mM ácido acético; 1 mMEDTA) a una temperatura de 60°C. La corrida electroforética se realizó a un voltaje constante de 160 V durante 3,5 h. Luego de la electroforesis, los geles fueron teñidos por 10 minutos en 250 mL búfer TAE 1X conteniendo 20 µL de una solución de 10 mg/ml de bromuro de etidio y desteñidos en 250 mL de búfer TAE 1X por 10 minutos. Finalmente los geles fueron fotografiados.

Capítulo 4

Resultados

4.1 Mejora de calidad de imagen desde la captura

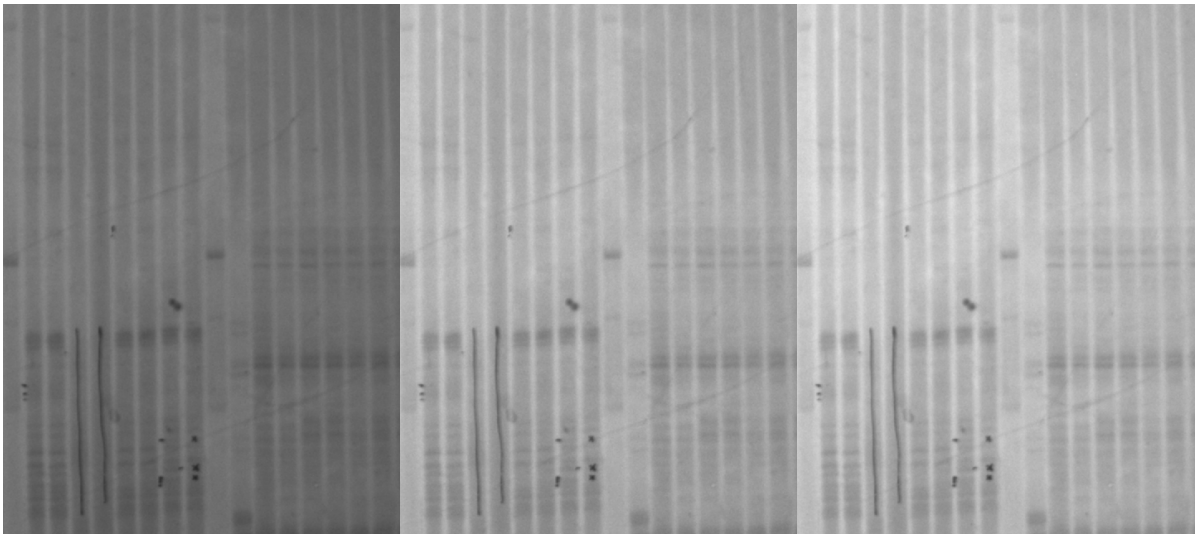
Los resultados detallados del sistema de captura adaptativo se presentan en [9]. Aquí se ilustran los resultados por el método de fusión de exposición utilizando los contrastes locales $\mathbf{C_{A_{lo}}}$ y $\mathbf{C_{RMS_{lo}}}$ (en los que las imágenes fusionadas presentaban los mayores valores de contraste y menor NRE) y la etapa de linealización para extender el rango dinámico de las imágenes.

Para la adquisición de imágenes se generan dos procesos de mejoramiento a diferentes niveles de iluminación: una en el día (iluminación natural) y otra en la noche (iluminación artificial) para comprobar que el sistema funciona y se adapta a distintos ambientes de prueba. Las imágenes adquiridas se muestran en la figura 4.1 para el día y en la figura 4.2 para la noche. La fusión y linealización de las imágenes capturadas se muestran en la figura 4.3 para el día y en la figura 4.4 para la noche.

En la tabla 4.1 se muestran las medidas globales de calidad para las imágenes obtenidas en el día y en la tabla 4.2 para la noche. Como medidas de aptitud locales se calculan los histogramas de intensidad local $L(i,j)$ y contraste absoluto local $\mathbf{C_{A_{lo}}}$ de las imágenes utilizadas en el proceso. Los histogramas para la prueba del día se muestran en la figura 4.5 y para la noche en la figura 4.6.

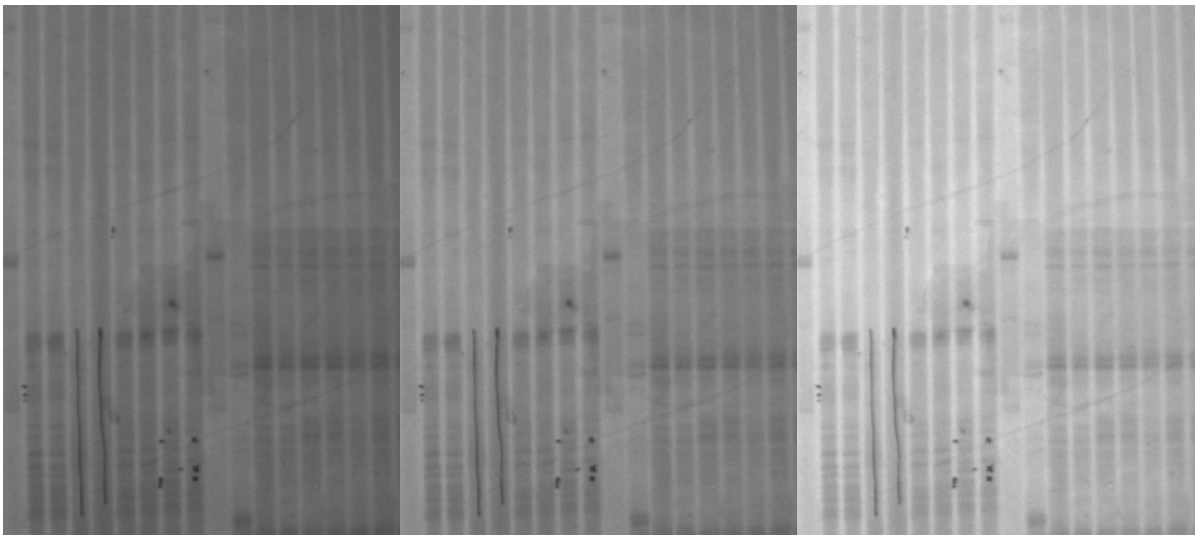
En los dos procesos evaluados (día y noche) se obtuvieron medidas de aptitud de intensidad μ_L y Me_L para las imágenes finales alrededor de 170 unidades de intensidad, como se muestra en tabla 4.1 para el día y en la tabla 4.2 para la noche. Además se observa que las medidas de aptitud para los dos ambientes de prueba presentan diferencias menores al 2% en las imágenes finales, lo que comprueba que el sistema funciona y se adapta a distintos ambientes de prueba.

En las tablas 4.1 y 4.2 se observa que el sistema de mejoramiento de imágenes digitales de geles de electroforesis por medio de fusión y ajuste multiparamétrico de la captura aumenta los valores de las cinco mediciones de contraste realizadas (absoluto, de Michel-



(a) I_o (b) I_m (c) I_c

Figura 4.1: Imágenes a fusionar adquiridas por el ajuste multiparamétrico en condiciones de de iluminación natural (día)

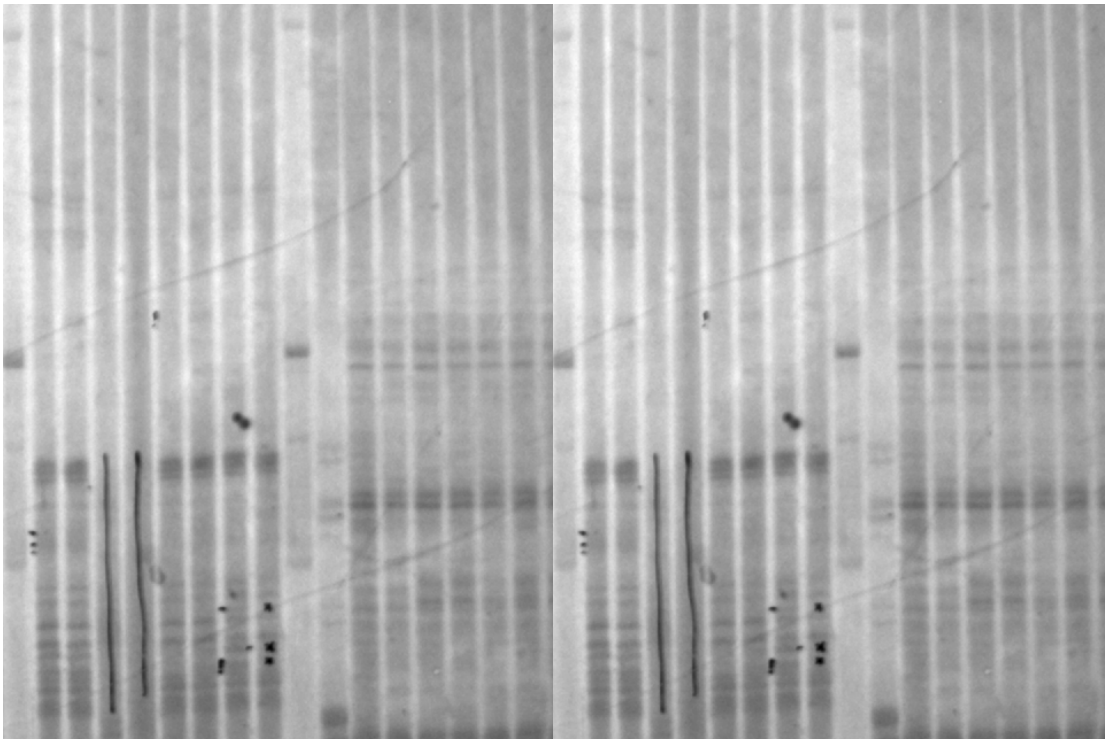


(a) I_o (b) I_m (c) I_c

Figura 4.2: Imágenes a fusionar adquiridas por el ajuste multiparamétrico en condiciones de iluminación artificial (noche)

son, de Weber, de intensidad y RMS) en sus versiones globales, alrededor de un 100% con respecto a la imagen con valores de contraste menores (imagen oscura) y en un 60% con respecto a la imagen con valores de contraste mayores (imagen clara). Además en las figuras 4.1 y 4.2 se observa que los histogramas de intensidad de las imágenes finales presentan un mayor rango dinámico, alrededor de un 35% más en la escala de grises que las imágenes de entrada. Así se comprueba el aumento de las medidas de aptitud de calidad de las imágenes capturadas.

Por otra parte, en las figuras 4.3 y 4.4 se observa que los histogramas de contraste absoluto



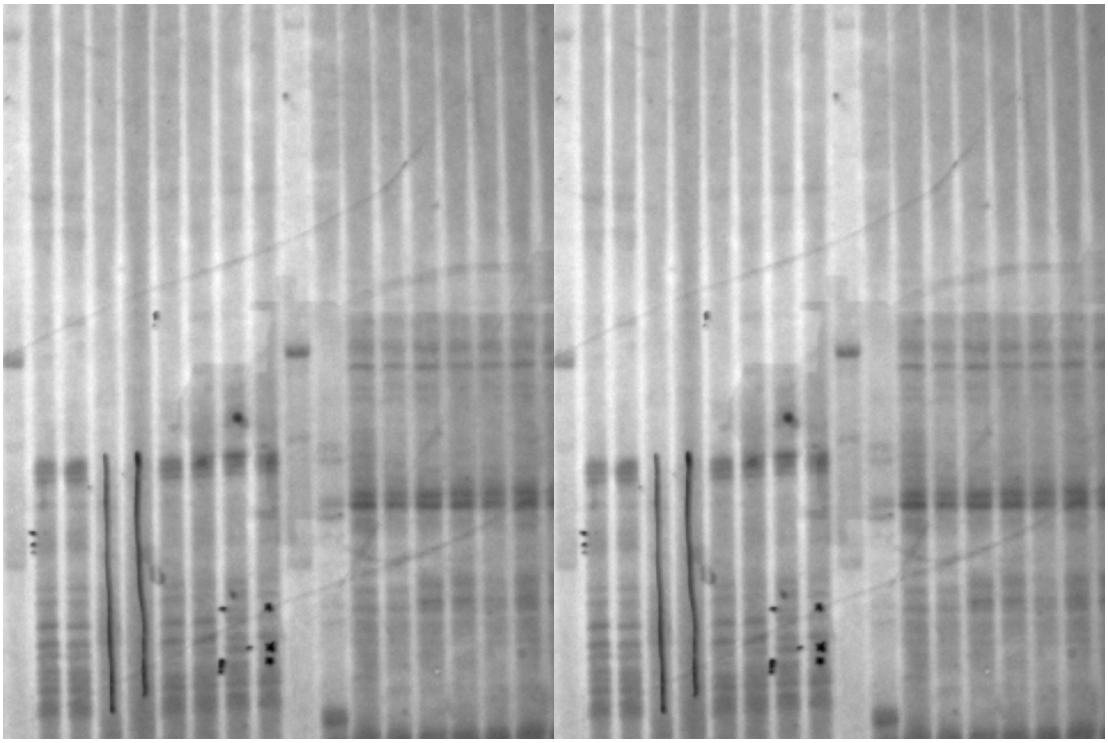
(a) $\mathbf{C_{A_{lo}}}$ (b) $\mathbf{C_{RMS_{lo}}}$

Figura 4.3: Imágenes finales del sistema diseñado en condiciones de iluminación natural (día)

Tabla 4.1: Resultados globales del mejoramiento de imágenes digitales de geles de electroforesis por medio de fusión y ajuste multiparamétrico de la captura en condiciones de iluminación natural (día)

	Imágenes iniciales Nivel de intensidad Figura 4.1			Imágenes finales EF y linealización Figura 4.3	
	4.1a	4.1b	4.1c	4.3a	4.3b
Parámetro	I_o	I_m	I_c	$\mathbf{C_{A_{lo}}}$	$\mathbf{C_{RMS_{lo}}}$
μ_L	111,489	168,116	184,817	178,7	178,815
Me_L	112	169	186	180	180
C_{A_G}	0,396	0,596	0,643	1	1
C_{M_G}	0,564	0,567	0,558	1	1
$C_{W_G min}$	-0,650	-0,655	-0,648	-1	-1
$C_{W_G max}$	0,255	0,249	0,239	0,427	0,426
C_{L_G}	0,905	0,904	0,887	1,427	1,426
C_{RMS_G}	11,162	16,288	17,809	29,01	29,02
NRE	4,759	8,146	8,802	16,44	17,6

local para las imágenes obtenidas por el sistema global presentan mayor rango dinámico que las imágenes de entrada, puesto que cubren alrededor de un 20% más de su rango total de valores y con valores más cercanos a su límite máximo. Ésto proporciona mayor



(a) $\mathbf{C_{A_{lo}}}$ (b) $\mathbf{C_{RMS_{lo}}}$

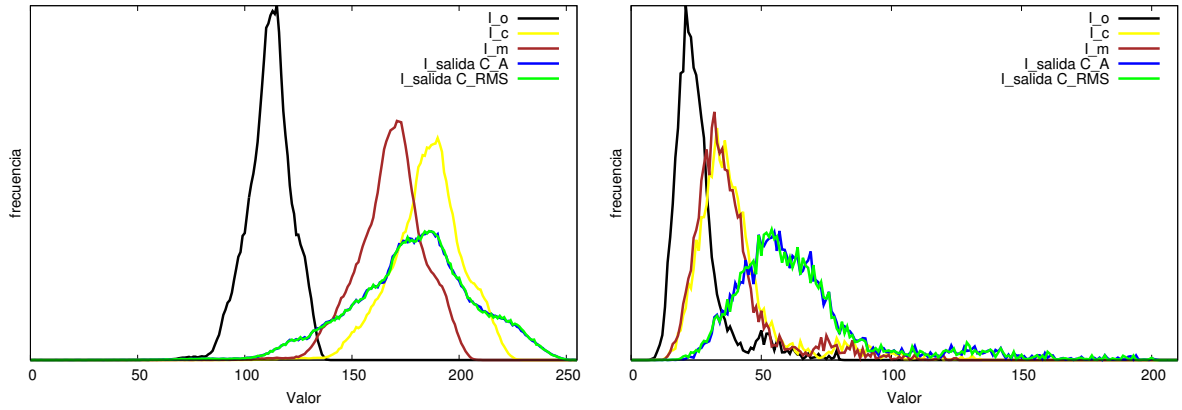
Figura 4.4: Imágenes finales del sistema diseñado en condiciones de iluminación artificial (noche)

Tabla 4.2: Resultados globales del mejoramiento de imágenes digitales de geles de electroforesis por medio de fusión y ajuste multiparamétrico de la captura en condiciones de iluminación artificial (noche)

	Imágenes iniciales Nivel de intensidad Figura 4.2			Imágenes finales EF y linealización Figura 4.4	
	4.2a	4.2b	4.2c	4.4a	4.4b
Parámetro	I_o	I_m	I_c	$\mathbf{C_{A_{lo}}}$	$\mathbf{C_{RMS_{lo}}}$
μ_L	110,932	133,971	184,368	172,756	173,367
Me_L	111	134	185	173	174
C_{A_G}	0,372	0,439	0,604	1	1
C_{M_G}	0,519	0,504	0,506	1	1
C_{W_Gmin}	-0,253	-0,589	-0,593	-1	-1
C_{W_Gmax}	0,253	0,246	0,242	0,476	0,471
C_{L_G}	0,856	0,836	0,835	1,476	1,471
C_{RMS_G}	10,616	11,85	17,014	30,778	31,027
NRE	4,319	5,765	8,245	17,64	16,95

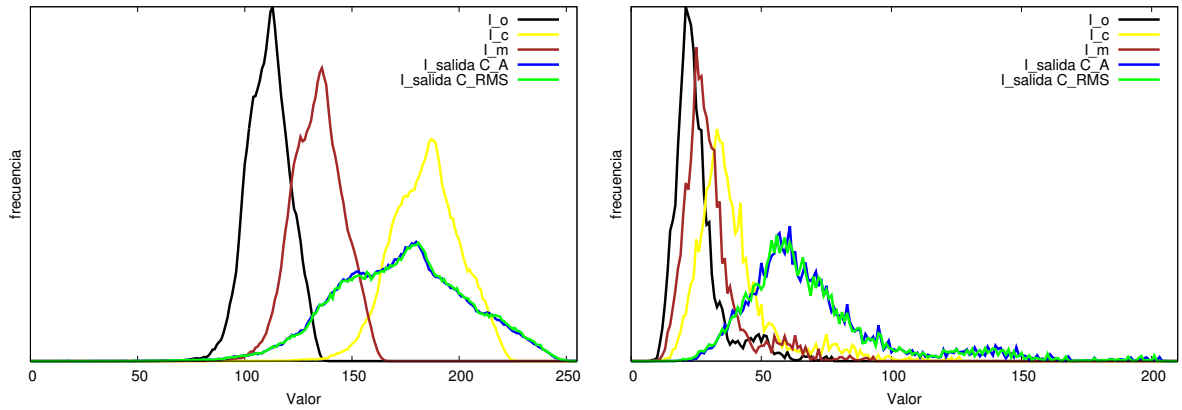
nitidez en las imágenes finales.

Por último en las tablas 4.1 y 4.2 se puede observar un aumento alrededor de 8 unidades



Intensidad local $L(i,j)$ Contraste $\mathbf{C_{A_{lo}}}$

Figura 4.5: Histogramas relevantes de las diferentes imágenes utilizadas en el proceso de mejoramiento de imágenes digitales de geles de electroforesis por medio de fusión y ajuste multiparamétrico de la captura en la prueba de día



Intensidad local $L(i,j)$ Contraste $\mathbf{C_{A_{lo}}}$

Figura 4.6: Histogramas relevantes de las diferentes imágenes utilizadas en el proceso de mejoramiento de imágenes digitales de geles de electroforesis por medio de fusión y ajuste multiparamétrico de la captura en la prueba de noche

de ruido en el NRE de las imágenes de entrada, debido a la fusión y la linealización realizadas. Este efecto se tolera debido a la disminución del 60% en las imágenes iniciales adquiridas por el sistema optimizador, por lo que el NRE en las imágenes finales adquiridas por el sistema global implementado son menores que las imágenes capturas sin éste en intensidades μ_L y Me_L aproximados, como las imágenes adquiridas de manera manual.

4.2 Detección de carriles y rectificación de imágenes

El análisis completo del método propuesto para la detección de carriles excede el marco de este informe, pero puede ser revisado en [13, 14]. A continuación se presentan algunos

resultados selectos.

4.2.1 Autocorrelación de las columnas del gradiente

Utilizando 50 puntos para la creación del ASM, se evaluó el comportamiento de la autocorrelación de las columnas número 1, 25 y 50 (figura 4.7) ante la suma de ruido blanco gaussiano a la totalidad la imagen (adicional al presente por naturaleza). Con este ruido se modelan indirectamente posibles degradaciones en la calidad de la imagen como reducción de contraste, presencia de ruido del CCD o difuminación de los carriles. Modificando

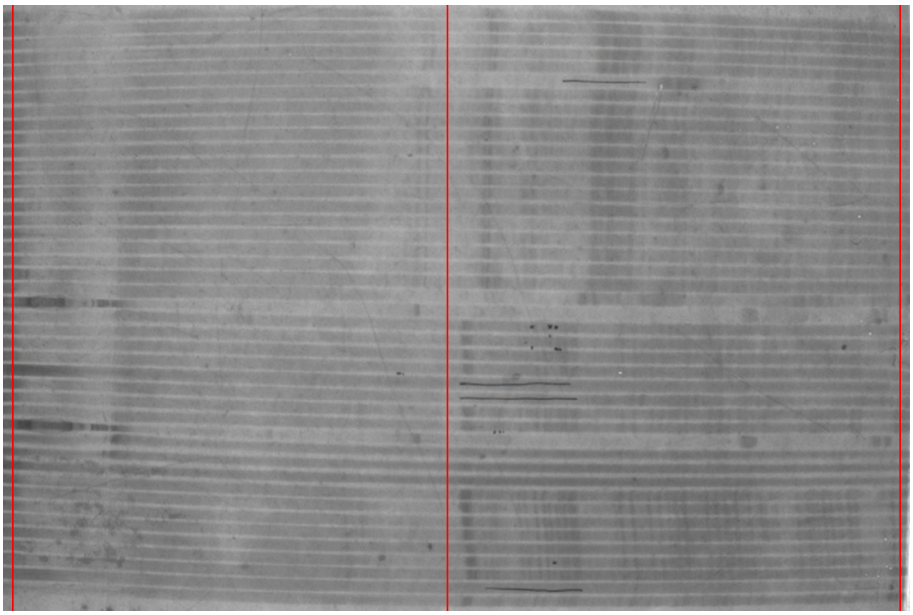


Figura 4.7: Columnas de prueba para la autocorrelación. De izquierda a derecha 1, 25 y 50

la varianza de la distribución de este ruido se logra aumentar o disminuir la amplitud que se suma. Representando esta amplitud como un porcentaje del máximo valor de la imagen se determina el ancho de los carriles de cada columna mediante la búsqueda del primer máximo a partir del origen. En la figura 4.8 se presenta la autocorrelación de dichas columnas sin ruido inducido para el tipo de gel que tiene separación entre carriles. El método de la autocorrelación, para la imagen con separación entre carriles, comienza a fallar para imágenes con separación entre carriles a partir de un 20% de ruido sumado a la imagen.

Para imágenes de geles sin separación entre carriles, se obtienen porcentajes de error más elevados respecto al caso anterior, que se justifican con la presencia de ruido en la imagen presente previo a las pruebas que provoca que no exista una diferencia de tonalidad de gris en la transición de un carril a otro.

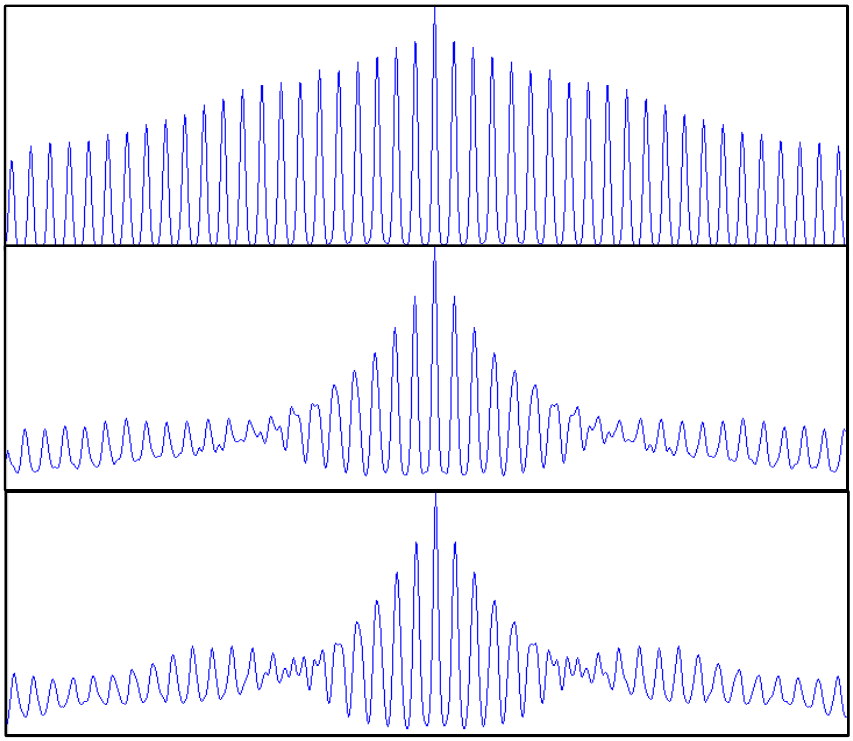


Figura 4.8: Correlación de las columnas. De arriba hacia abajo corresponden a la número 1, 25 y 50

4.2.2 Detección de carriles

Imágenes con separación entre carriles

En la figura 4.9 se muestra la detección de carriles en una sección de una imagen de prueba utilizando el algoritmo propuesto, 50 puntos para describir cada forma y un 95% de la varianza total del conjunto de entrenamiento considerada. En ésta las líneas azules indican los bordes de los carriles.

Para evaluar la detección se marcan los carriles manualmente sobre la imagen de prueba. Hecho esto se compara la detección manual con la obtenida con el algoritmo propuesto. Sea M el número de formas, N la cantidad de hitos en cada una de ellas, p_{iy} la coordenada y del punto detectado por el criterio humano y j_{iy} la detectada por el computador el los dos criterios de evaluación se evalúa la diferencia utilizando:

$$D = \frac{1}{MN} \sum_{i=0}^{MN-1} |p_{iy} - j_{iy}|$$

Se mide la detección de carriles en función del porcentaje de la varianza total del conjunto de entrenamiento utilizada en la creación del ASM, manteniendo la cantidad de puntos por forma en 110. Dicho porcentaje corresponde a la suma de los valores propios correspondientes a las dimensiones utilizadas del modelo de forma, dividida entre la suma de todos los valores propios de la descomposición de componentes principales. De acuerdo

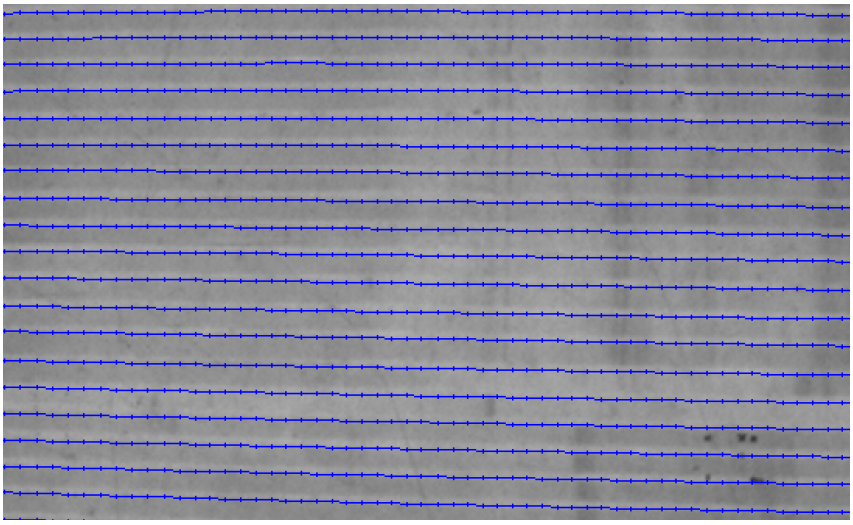


Figura 4.9: Detección de carriles con el algoritmo implementado

Tabla 4.3: Desviación en la detección de carriles en función del porcentaje del total de la varianza utilizada en el entrenamiento

% varianza total/dimensiones	D (píxeles)
50/1	4,2512
80/2	4,3161
96/3	4,2836
100/220	4,2537

con la tabla 4.3 la detección de los carriles no es afectada por el número de dimensiones consideradas; sin embargo, como se verá en la siguiente sección, esto no implica que se dé una mejor corrección de las distorsiones.

Imágenes sin separación entre carriles

En la figura 4.10 se muestra la detección de carriles utilizando el algoritmo para una sección de una imagen de prueba sin separación intercarril.

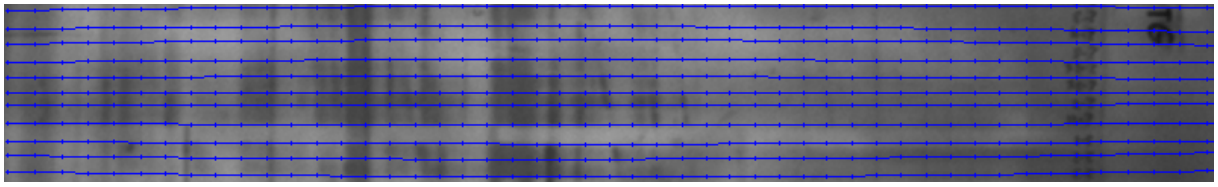


Figura 4.10: Detección de carriles para una sección de imagen de geles sin separación intercarril.

En la tabla 4.4 se observan los resultados de la desviación en función de la cantidad de dimensiones utilizadas. Para el caso se utilizan formas con 50 puntos. Se observa que la

Tabla 4.4: Desviación en la detección de carriles en función del porcentaje del total de la varianza utilizada en el entrenamiento

% varianza total/dimensiones	D (píxeles)
50/1	3,0716
80/2	3,0381
96/3	3,0596
100/100	3,1149

desviación promedio en la detección de carriles presenta poca variación. Sin embargo, se verá en la siguiente sección que no implica una buena corrección de las distorsiones.

4.2.3 Rectificación de la imagen

Imágenes con separación entre carriles

El proceso iterativo de ajuste y el del mapeo que rectifica la imagen se evalúan en esta sección obteniendo la desviación estándar de tres distintos carriles (etiquetados como 1, 2 y 3) antes y luego de la rectificación. Los valores antes de la rectificación de la imagen ejemplo se resumen en la figura 4.5 y figura 4.6.

Tabla 4.5: Desviación estándar inicial de los carriles antes de la rectificación

Carril	Desv. Std. [px]
1	8,2797
2	1,8139
3	8,4060

Se evalúa variando la cantidad de puntos por forma, obteniendo las desviaciones estándar de los carriles antes y después de la rectificación.

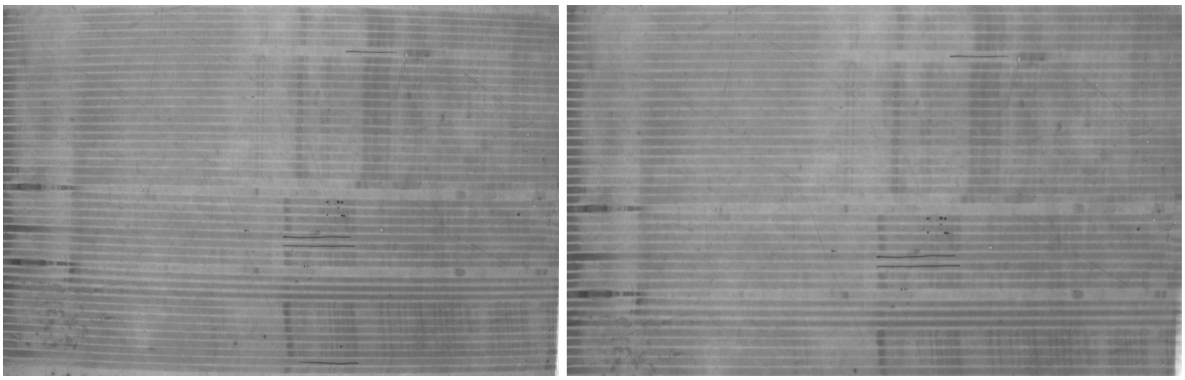
Tabla 4.6: Desviación estándar de los carriles luego de la rectificación en función del número de puntos por forma

Pts ASM	Desv. Std. [px]		
	Carril 1	Carril 2	Carril 2
20	1,0601	1,2546	1,6511
50	1,1634	1,2610	1,6597
80	0,7438	0,7529	1,5387
110	0,8046	0,7996	0,7407
140	0,7641	0,9994	0,9241

En ningún caso las desviaciones estándar fueron más altas que la inicial, lo que indica que siempre se redujeron las distorsiones. El hecho de que la desviación no se acercara a cero se debe a factores como distorsiones en los carriles que no son consideradas en el entrenamiento del modelo o imperfecciones en el proceso de creación de los geles como falta de uniformidad en el ancho de los carriles.

Para los tres carriles se nota una tendencia a aumentar la corrección conforme se aumenta el número de puntos. Esto se debe a que al haber más puntos se describe la curvatura que presentan los carriles con mayor precisión.

El conjunto más bajo de desviaciones estándar se obtuvo para 110 puntos. El resultado de la imagen rectificada para este caso se muestra en la figura 4.11.



(a) (b)

Figura 4.11: Imagen original (a) y rectificada (b).

Se evalúa la corrección modificando el porcentaje de la varianza total del conjunto de entrenamiento que es considerada para determinar la cantidad de dimensiones utilizadas por el ASM. En este caso se fija el número de puntos por forma en 110. Los resultados se muestran en la tabla 4.7.

Tabla 4.7: Desviación estándar de los carriles luego de la rectificación en función del número de puntos por forma

% varianza total/dimensiones	**Desv. Std. [px]**		
	Carril 1	Carril 2	Carril 2
50/1	3,8114	0,5376	2,2317
80/2	0,9851	0,9049	0,5919
96/3	0,9163	0,8980	0,5580
100/100	0,7437	1,2336	1,5041

Se comprueba que el problema no puede ser representable en una sola dimensión y que un mínimo de dos es necesario. Se debe destacar que al usar todas las dimensiones se da al modelo la libertad de variar siguiendo cualquier forma del conjunto de entrenamiento, permitiendo que una forma adopte distorsiones que no están dentro del rango considerado.

El detalle de los resultados del proceso de rectificación se presenta en [13].

4.3 Estimación del efecto sonrisa

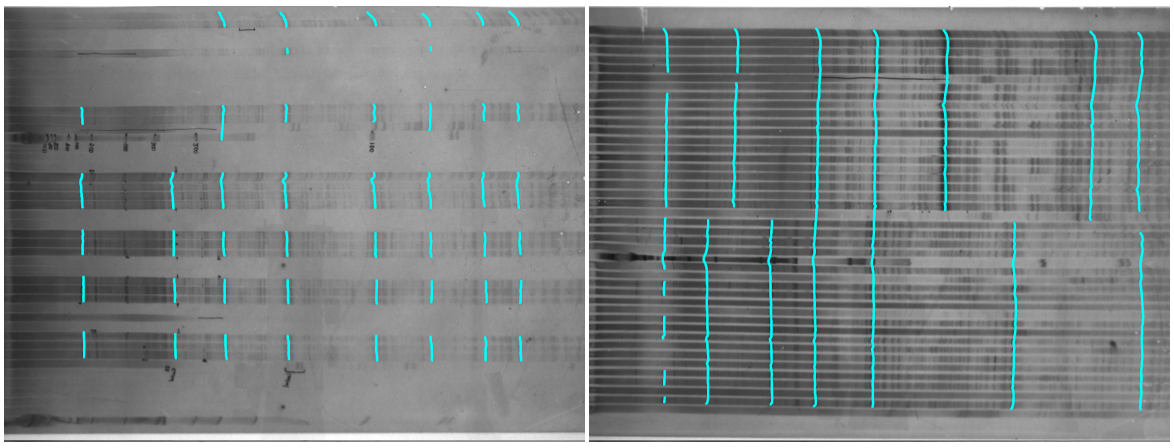
El detalle de los resultados obtenidos con la estimación del efecto sonrisa, así como su evaluación por medio de frentes de Pareto se presenta en [5, 4]. Se presentan a continuación algunas de las estrategias novedosas de evaluación del efecto sonrisa desarrolladas para el proyecto.

4.3.1 Funciones de aptitud

Para evaluar el sistema implementado se utilizan como funciones de aptitud el rendimiento, la corrección del efecto sonrisa y confianza de las formas, las cuales se detallan a continuación.

- Rendimiento: Esta función evalúa la velocidad con la es aplicado el algoritmo. Es medido en imágenes procesadas por minuto.

- Corrección del efecto sonrisa: Para determinar la corrección realizada sobre la imagen se señala sobre el grupo de imágenes de prueba de manera manual las líneas de bandas presentes en ellas como se muestra en la figura 4.12.



(a) (b)

Figura 4.12: Imágenes del grupo de pruebas con líneas de bandas marcadas

Como se observa en la figuras 4.12 solo son señaladas las líneas de bandas donde se puede determinar la existencia de bandas, esto para evitar que al suponer la posición de las bandas se altere la distorsión presente en la imagen y por lo tanto los resultados.

Los segmentos marcados son leídos y unidos en líneas de bandas (en el caso de que pertenezcan a la misma línea de banda) pero para el calculo de la distorsión presente sólo son tomados en cuenta los elementos marcados de manera manual.

Se mide el efecto sonrisa con la desviación estándar de las líneas marcadas de manera manual. La corrección del efecto sonrisa es medido como la reducción de esa desviación estándar. Esto se obtiene mediante

$$\Delta\sigma = \frac{\sigma_I - \sigma_C}{\sigma_I} \tag{4.1}$$

donde σ_I es la desviación estándar presente en la imagen a corregir y σ_C es la desviación estándar presente en la imagen luego de pasar por el sistema de corrección de efecto sonrisa. La razón $\Delta\sigma$ se puede convertir en el porcentaje de la reducción del efecto sonrisa si es multiplicada por 100%.

- Confianza de las formas: Esta prueba indica la suma de la confianza que poseen todos los elementos de las formas. Este criterio además de medir qué tan fuertes son los bordes en los que se encuentra ubicada las formas, también mide de manera indirecta la cantidad de formas creadas para corregir la imagen. Una gran cantidad de formas presentes en la imagen no es algo deseable ya que existe la posibilidad de que dichas formas se crucen y produzcan una mayor distorsión que la ya presente en la imagen

4.3.2 Evaluación multiobjetivo con frentes de Pareto

La evaluación genética empleada permite optimizar el desempeño del algoritmo en todas las parametrizaciones, buscando mejorar en forma simultánea todas las medidas de aptitud elegidas.

El algoritmo genético utilizado es el PESA [8] que se encuentra incluido en la LTI-Lib2 [20]. Este algoritmo realiza la variación de los parámetros mediante mutaciones y cruces entre grupos de ellos para intentar obtener un mejor resultado en alguna de las funciones de aptitud que se utilizan.

En la figura 4.13 se muestra el frente de Pareto obtenido utilizando la primera imagen del banco de pruebas como sujeto. En ella se puede observar que valores altos en la corrección del efecto sonrisa y confianza de las formas implican un menor rendimiento. Además se observa que la corrección del efecto sonrisa decrece cuando se aumenta la confianza de las formas.

Del frente de Pareto mostrado en la figura 4.13 se extraen los valores máximos para cada una de las funciones de aptitud, los parámetros para obtener dichos valores se muestran en la tabla 4.8.

Para la primera imagen del banco de pruebas el sistema diseñado logra reducir en un 99.736% la distorsión efecto sonrisa presente en la imagen, procesando 3.2156 imágenes

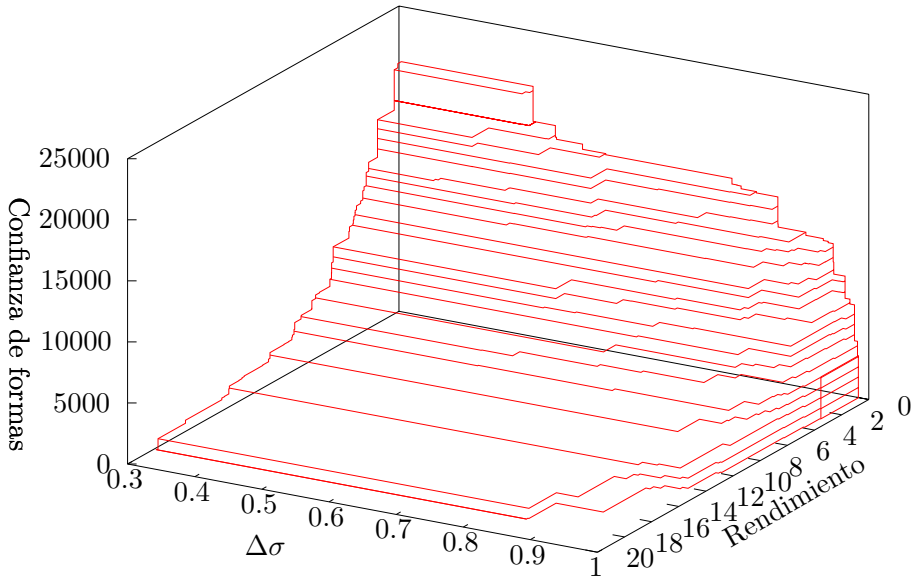


Figura 4.13: Frente de Pareto de pruebas realizadas sobre la primera imagen

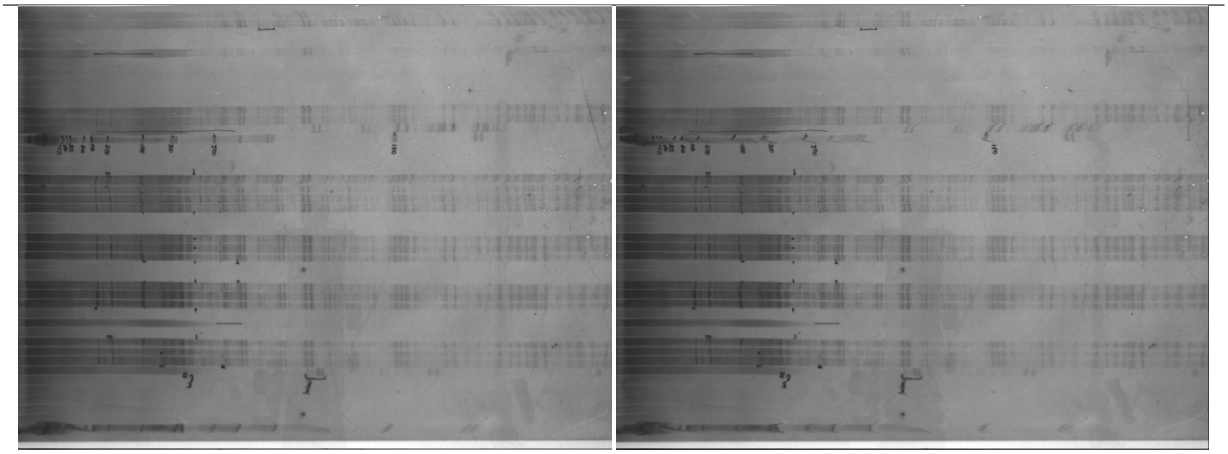
Tabla 4.8: Parámetros para valores máximos de funciones de aptitud para la primera imagen

Parámetros	**Max. Rendimiento**	**Max. Confianza**	**Max. Corrección**
Rendimiento	18.4893	0.704144	3.21956
Confianza	378.443	20859	3809.89
Corrección	0.863752	0.509193	0.99376
maxUm	0.091753	0.1074	0.110529
smootK	1	1	2.80645
hessianK	5	5	7.90323
dT	0.905976	0.404518	0.404518
T	1.30588	39.2353	3.75294
α	0.001737	0.003218	0.002361
C	1×10^{-15}	1×10^{-15}	8×10^{-16}
confMascX	11	9	12
confMascY	14	11	4
maxVecin	3	7	2
maxDifFu	10	3	12
nMaxRect	10	9	21
maxLPConf	15	16	160

por minuto. El sistema puede procesar máximo 18.4893 de las primeras imágenes por minuto reduciendo en ellas la distorsión del efecto sonrisa en un 86.3752%. Además la máxima confianza de las formas que se obtiene para dicha imagen es de 20859 corrigiendo la imagen en un 50.9193%, procesándolas a una velocidad de 0.704144 imágenes por minuto.

Las figuras 4.14 y 4.15 ilustran el resultado obtenido para la mayor corrección del efecto sonrisa mediante la imagen original y las líneas de bandas señaladas de manera manual

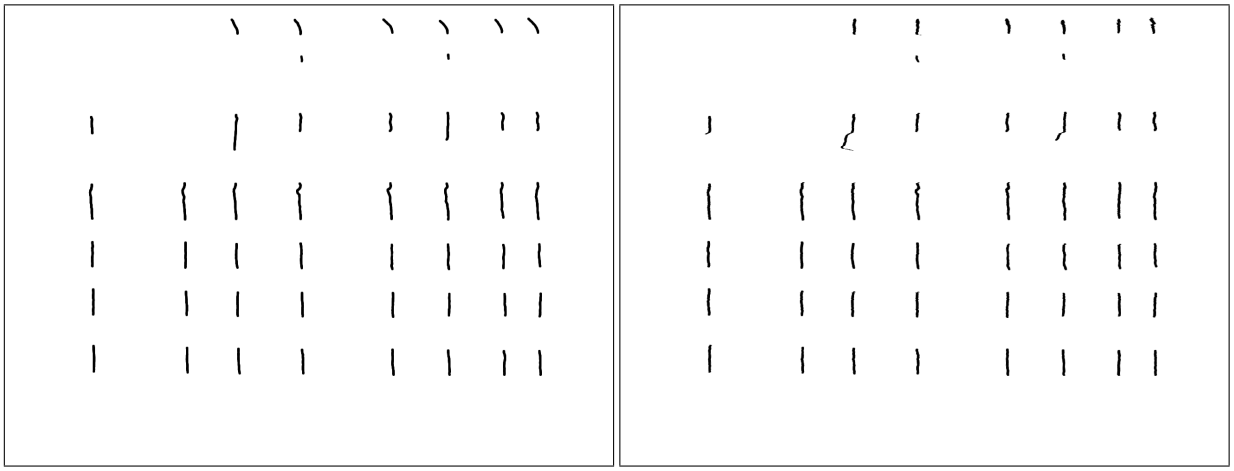
respectivamente.



(a) (b)

Figura 4.14: Primera imagen del banco de pruebas antes (a) y después de ser procesada (b)

El la figura 4.14 se puede observar como el rayón presente al lado derecho de la imagen fue completamente rectificado, esto implica que los modelos de formas se adaptaron al rayón y no a las líneas de bandas cercanas a este, lo que hace que la corrección del efecto sonrisa no sea la máxima que se puede obtener en una imagen de ese tipo.



(a) (b)

Figura 4.15: Líneas de bandas marcadas para la primera imagen del banco de pruebas antes (superior) y después de ser procesada (inferior)

El procedimiento anterior se repite para la segunda imagen, y de nuevo para las dos imagenes de forma conjunta (ver detalles en [5]), de modo que se pueda evaluar la estabilididad del algoritmo ante los parámetros. Como es de esperar, la optimización conjunta alcanza un rendimiento total menor que en forma aislada, como en el caso expuesto anteriormente, pero aun así se logra reducir considerablemente la distorsión en cuestión.

Para elevar el porcentaje de la corrección del efecto sonrisa se debe tener una cantidad de formas que reflejen la distorsión real de la imagen y como se mencionó anteriormente

una mayor cantidad de formas requiere un mayor tiempo de CED. Por lo tanto obtener un porcentaje mayor de corrección implica un menor rendimiento.

Se confirma que la solución diseñada reduce su capacidad de corregir el efecto sonrisa si la confianza en las formas crece. Esto se debe a que para tener una mayor confianza en las formas se debe tener mayor cantidad de ellas. Las posibilidades de cruces entre formas se aumentan cuando su cantidad es mayor. Por lo tanto una mayor cantidad de formas puede introducir distorsión en la imagen en lugar de reducirla.

Los valores de dT y de T son con los que controlan la cantidad de iteraciones que se utilizan para la CED de forma que

$$\text{número de iteraciones} = \frac{T}{dT} \tag{4.2}$$

Una menor cantidad de iteraciones produce un mejor rendimiento del sistema. Por lo tanto cuanto menor el valor de T y mayor el valor dT mejor será el rendimiento del sistema.

El valor más alto de corrección del efecto sonrisa obtenido en las pruebas es de un 99.376%, el cual se consigue para la primera imagen del banco de pruebas.

También es posible determinar que el sistema y sus parámetros son dependientes de las imágenes que se utilicen, esto quiere decir que el sistema con un conjunto de parámetros puede corregir casi en un 100% la distorsión del efecto sonrisa de una imagen pero el mismo conjunto de parámetros no necesariamente obtendrá un valor alto de corrección en otra imagen. Esto esta demostrado ya que cuando se analizó ambas imágenes por separado se obtuvieron valores de corrección superiores al 99% pero al analizarlas juntas nunca se superó el 50%.

4.4 Detección de bandas

4.4.1 Detección por medio de optimización

Para la medición cuantitativa de los algoritmos de detección de bandas se utilizan tanto carriles reales extraídos de geles, como carriles sintéticos, es decir, carriles creados algorítmicamente con una cantidad de bandas establecida manualmente, distribuidas aleatoriamente a lo largo del carril. De igual forma sucede con la intensidad de las bandas, las cuales son generadas aleatoriamente dentro de un rango establecido. Para estos carriles sintéticos se conoce el vector de parámetros que dio origen al carril, permitiendo así realizar evaluaciones del rendimiento del método propuesto, considerando la diferencia entre la cantidad de bandas en el carril y la cantidad de bandas estimadas, así como la desviación promedio entre la ubicación de las bandas originales y la ubicación estimada y el error total existente entre la distribución de intensidad estimada y la real.

El funcionamiento del optimizador de la función objetivo se evalúa con la ventana de un

carril sintético, mostrada en la figura 4.16, de tamaño 100×20 con 6 bandas caracterizadas por un perfil de intensidad con $\sigma = 1\,\text{px}$.

Figura 4.16: Carril sintético para la evaluaciíon de la optimización de la función objetivo.

En la tesis [59] se presenta una evaluación detallada del algoritmo para varios casos de estudio así como para cada fase de la propuesta. La figura 4.17 presenta como ejemplo un segmento de carril real, junto a su estimación así como el perfil real y el estimado.

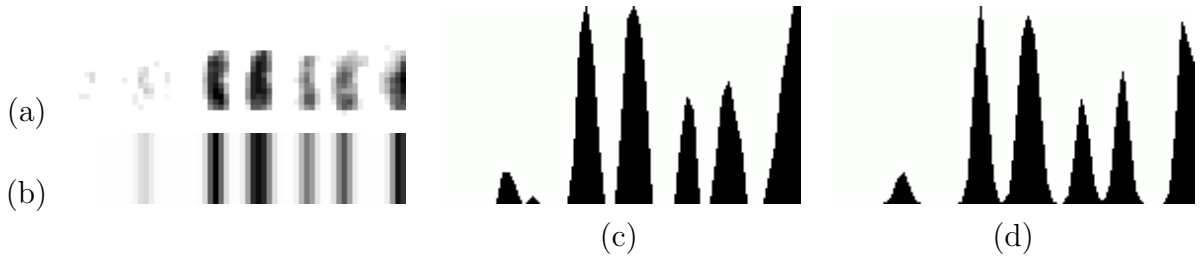


Figura 4.17: Ejemplo de detección de bandas para un (a) carril real (b) carril estimado (c) perfil real, (d) perfil estimado.

Si bien es cierto el método por optimización es robusto, su complejidad computacional es elevada, requiriendo hasta 20 s para realizar la estimación de un único carril, sin contemplar la estimación del ancho de las bandas σ.

4.4.2 Incorporación de estimación del ancho de las bandas

El detalle de resultados del algoritmo completo se presenta en [22]. La localización de las bandas en un carril de electroforesis puede ser realizada utilizando la segunda derivada, debido a que esta aumenta los cambios en la señal [25], por lo cual facilita la tarea de búsqueda de máximos.

El aumento de la escala deja una señal cada vez con menos estructura, tal como se muestra en la Figura 4.18, donde se muestra cómo al aumentar la escala las bandas cercanas se unen en el espacio de escalas formando una sola banda y un único máximo, por lo que la segunda derivada provee la mejor resolución para encontrar las bandas, la cual corresponde a la fila inicial del espacio de escalas.

En [31] se destaca que para la caracterización de las bandas es más conveniente utilizar la más alta resolución del espacio de escalas, que corresponde a la segunda derivada del carril, porque al aumentar la escala los cruces por cero de la derivada y los máximos se desplazan más significativamente. En un caso sin traslape los cruces por cero de la segunda derivada se encuentran situados en $\pm\sigma$ a partir del punto central de la banda y la identificación de estos permitiría determinar la desviación estándar en forma directa.

Figura 4.18: Espacio de escalas con traslape de bandas.

La derivada realizada en este trabajo se hace utilizando el método de filtrado y derivación se Savitzky y Golay [53] para encontrar directamente la segunda derivada.

Aunque las posiciones correspondientes a los máximos de la segunda derivada y los cruces por cero no correspondan exactamente con los valores teóricos aun considerando funciones continuas, estos datos aproximan lo suficiente a las bandas como para estimar la cantidad de bandas en los datos considerados [31].

Para evaluar el comportamiento de este método ante el traslape se utiliza un carril con 26 bandas con desviación estándar $\sigma = 5$, el cual se muestra en la figura 4.19.

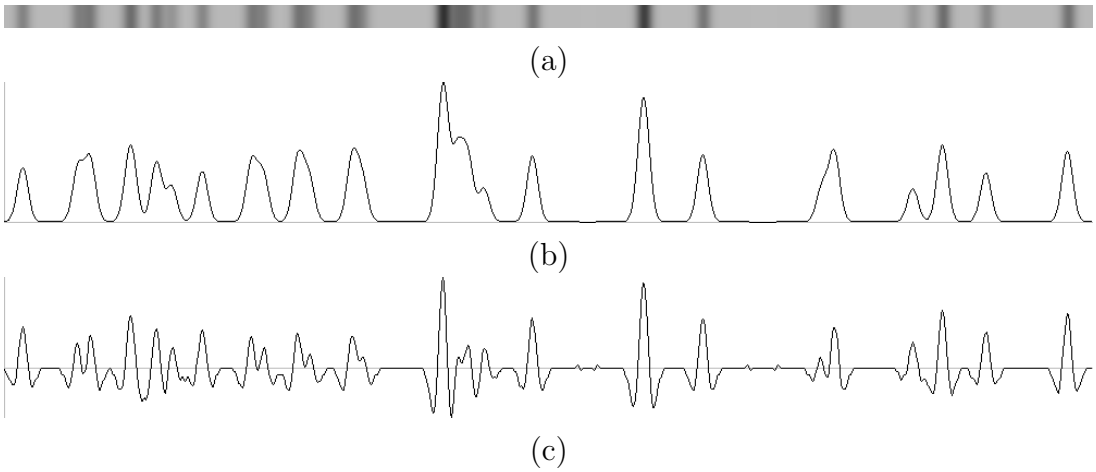


Figura 4.19: Carril teórico con mayor traslape. (a) Perfil del carril, (b) perfil invertido y (c) su segunda derivada.

Esta imagen muestra el efecto de mejora de la resolución de los picos en el carril, lo cual permite distinguir aún aquellos que a la vista parecen como una única banda más ancha

producto de la suma de dos bandas adyacentes.

Las posiciones de cada una de estas bandas se muestra en la tabla 4.9, en la cual se muestran las posiciones con respecto a sus valores teóricos, así como el error en píxeles en que se incurre en la determinación de la posición de cada una de estas bandas, que en el peor de los casos llega a ser de hasta 3 píxeles, esto es, 0,6 veces la desviación estándar de la banda.

Tabla 4.9: Posición de las bandas en un carril sintético con 26 bandas

Posición teórica	Posición obtenida	Error (píxeles)	Posición teórica	Posición obtenida	Error (píxeles)
17	17	0	403	403	0
70	67	3	416	418	2
78	79	1	423	426	3
116	116	0	440	441	1
140	140	0	485	485	0
154	155	1	587	587	0
182	182	0	642	642	0
228	227	1	753	750	3
237	239	2	762	763	1
271	269	2	835	835	0
279	281	2	862	862	0
321	320	1	902	902	0
327	329	2	977	977	0

Determinación de la cantidad de bandas en el carril

La segunda derivada permite distinguir entre bandas muy cercanas entre sí, tal como se ejemplificó en la sección anterior en el caso de carriles con traslape de bandas. Con el objetivo de analizar la determinación de la cantidad de las bandas que el algoritmo es capaz de detectar, se generan 10 carriles sintéticos para medir la cantidad de bandas detectadas para cada valor de desviación estándar empezando en 1 hasta 5,5 con paso de 0,5, cada uno con 20 bandas.

En estas mediciones interesa evaluar la detección de bandas en comparación con lo que un usuario a simple vista podría realizar, por lo tanto se genera la tabla 4.10 con 100 mediciones donde se realizó un conteo de las bandas detectables a simple vista, y en la tabla 4.10 se muestran los resultados de la detección de bandas utilizando la segunda derivada.

La cantidad de bandas en estos carriles es teóricamente de 20, sin embargo, el algoritmo generador de carriles sintéticos coloca algunas de las bandas en posiciones tan cercanas en algunos casos que ni siquiera con la segunda derivada se puede distinguir entre dos bandas adyacentes.

Tabla 4.10: Evaluación de cantidad de bandas en carriles sintéticos. Cantidad de bandas observadas.

σ \ Medición	1	2	3	4	5	6	7	8	9	10
1	20	20	19	18	19	20	20	18	20	20
1.5	19	18	19	18	17	15	18	20	17	19
2	17	16	20	17	19	19	17	20	20	19
2.5	15	18	15	17	17	16	15	17	19	19
3	15	16	18	16	16	17	15	17	14	16
3.5	14	16	14	15	17	18	18	14	14	16
4	14	19	17	15	17	20	18	18	17	17
4.5	18	18	15	19	18	19	17	16	18	16
5	16	17	16	17	13	17	16	15	16	16
5.5	18	15	16	16	18	15	14	14	17	15

Tabla 4.11: Evaluación de cantidad de bandas en carriles sintéticos. Cantidad de bandas obtenidas mediante la segunda derivada.

σ \ Medición	1	2	3	4	5	6	7	8	9	10
1	20	20	19	18	19	20	20	18	20	20
1.5	19	18	19	18	17	16	18	20	17	19
2	17	16	20	17	19	19	17	20	20	19
2.5	16	18	15	17	18	16	15	19	19	19
3	16	17	18	17	16	17	16	18	16	16
3.5	15	16	14	15	17	18	19	15	15	16
4	15	20	17	16	18	20	19	18	17	18
4.5	18	19	15	20	18	19	17	16	18	16
5	17	18	16	18	13	17	16	15	17	16
5.5	19	15	16	18	19	16	15	15	17	16

La cantidad de bandas detectadas mediante la segunda derivada mejora la resolución en varios de los casos analizados, puesto que para ninguno de los 100 carriles sintéticos analizados la cantidad de bandas estimadas a simple vista fue mayor a la cantidad estimada por el algoritmo. El aumento en la cantidad de bandas detectadas obedece a la mejora de resolución aplicada en los casos en que las bandas se traslapan y que a simple vista puede parecer como una única banda más ancha en comparación con las demás, por lo tanto se considera satisfecho el criterio de identificación de más del 95% de las bandas del carril con respecto a la cantidad de bandas observadas, dado que es imposible determinar en este caso la totalidad de las 20 bandas a menos que no estén muy traslapadas.

Determinación de la desviación estándar con Line Search y espacio de escalas

La localización de las bandas por medio de la segunda derivada permite únicamente conocer la ubicación; sin embargo es necesario obtener el valor correspondiente a la desviación estándar considerando todas las bandas encontradas. En este trabajo se asume que las bandas de los carriles poseen el mismo ancho.

Una vez conocidas las ubicaciones de las diferentes bandas, se realiza la optimización de la función objetivo, centrando las bandas en los máximos detectados, de forma que el error obtenido sea el mínimo al realizar el ajuste para diferentes valores de desviación estándar. El valor inicial de desviación estándar utilizado para aproximar estas bandas es el resultado de la aplicación del espacio de escalas.

El ajuste realiza la minimización del error en la aproximación mediante mínimos cuadrados utilizando el optimizador *LineSearch* [53] en una dimensión. En la tabla 4.12 se encuentran los resultados correspondientes a la aplicación de la optimización de la función objetivo considerando carriles sintéticos con una banda.

Tabla 4.12: Estimación de desviación estándar utilizando la función objetivo para carriles teóricos con una banda

σ teórico	σ promedio	C.V. (%)	Error (%) promedio	Error máximo 1 píxel (%)
0,5	0,578	16,239	15,597	33,333
1	1,043	2,661	4,270	16,667
1,5	1,548	1,983	3,230	11,111
2	2,041	0,981	2,055	8,333
2,5	2,537	0,574	1,471	6,667
3	3,034	0,478	1,128	5,556
3,5	3,535	0,430	0,995	4,762
4	4,029	0,384	0,736	4,167
4,5	4,535	0,411	0,779	3,704
5	5,045	0,432	0,895	3,333
5,5	5,537	0,355	0,667	3,030
6	6,036	0,531	0,596	2,778
6,5	6,564	0,456	0,987	2,564
7	7,037	0,534	0,522	2,381
7,5	7,540	0,465	0,528	2,222
8	8,030	0,640	0,374	2,083
8,5	8,516	0,315	0,183	1,961
9	9,027	0,378	0,303	1,852
9,5	9,530	0,233	0,314	1,754
10	10,051	0,635	0,506	1,667

Para cada valor de desviación estándar se realiza el cálculo utilizando 10 carriles sintéticos

y se obtienen los valores medios de la desviación estándar y del error con respecto al valor teórico.

El error obtenido para las bandas con menor desviación estándar se reduce en comparación a cuando solo se usa el espacio de escalas como se muestra en la figura 4.20, donde destaca que para $\sigma = 0,5$ el error pasa de 113,20% a 15,597%. Además se incluye el error máximo permitido en la última columna, el cual es de 1 píxel para cada una de las bandas, donde se nota que para el método propuesto las bandas pueden ser aproximadas con error menor a este límite. Este error máximo está determinado por la siguiente ecuación

$$e_{max} = \frac{1}{6\sigma} 100\% \tag{4.3}$$

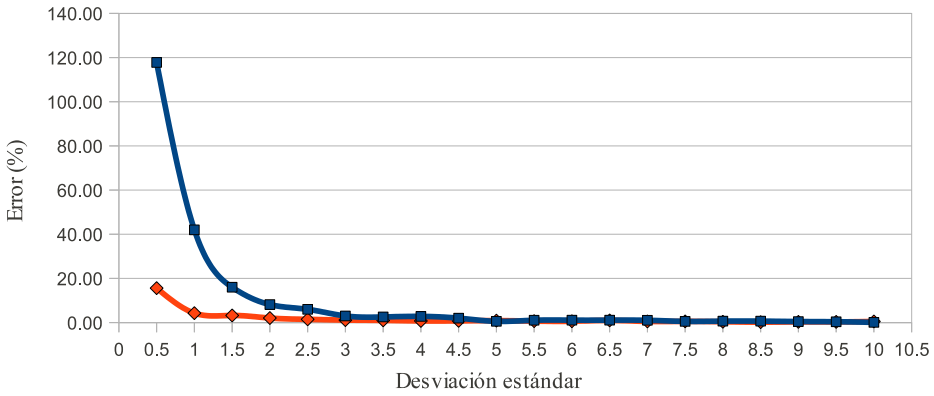


Figura 4.20: Error en la determinación de la desviación estándar en carriles con una banda. Error con espacio de escalas (□), error con espacio de escalas y mínimos cuadrados (◇) .

La figura 4.20 muestra la mejora en la aproximación de la desviación estándar de las bandas utilizando el espacio de escalas en conjunto con el método de mínimos cuadrados en comparación con la utilización simple del espacio de escalas para carriles con una única banda. En este gráfico sí se incluye el error para las bandas con $\sigma = 0,5$ puesto que aunque para el espacio de escalas este error es de $113,2\%$, permite realizar la comparación con el error generado con el método propuesto para este ancho de la banda.

En la tabla 4.13 se muestran los resultados para la evaluación del algoritmo considerando desviaciones estándar de 0,5 a 10, con paso de 0,5, para 10 carriles con 20 bandas para cada valor de desviación estándar. Los carriles generados tienen la restricción de una separación mínima entre bandas de 2σ, de forma que todas las bandas puedan ser distinguidas.

La desviación estándar está dentro de los límites aceptables en este caso, puesto que todos los errores obtenidos son menores a un píxel, y los coeficientes de variación son menores a 3.277%, lo cual indica que los datos son precisos y estables utilizando este algoritmo, aún si las bandas se encuentran traslapadas.

Tabla 4.13: Evaluación de la desviación estándar utilizando carriles sintéticos con 20 bandas traslapadas. Aproximación del espacio de escalas y minimización lineal.

σ teórico	σ Promedio	C.V. (%)	Error (%)	σ teórico	σ Promedio	C.V. (%)	Error (%)
0,5	0,645	3,277	20,137	5,5	5,459	0,907	1,973
1	1,149	3,490	8,052	6	5,974	0,774	0,160
1,5	1,526	3,369	3,473	6,5	6,440	0,791	0,226
2	2,018	1,426	0,310	7	6,967	1,032	0,350
2,5	2,515	0,883	2,576	7,5	7,477	0,384	0,567
3	3,003	0,583	0,042	8	7,990	0,473	0,518
3,5	3,523	0,298	0,611	8,5	8,480	0,436	0,350
4	3,978	0,767	0,484	9	8,986	0,467	0,309
4,5	4,484	1,432	0,330	9,5	9,476	0,549	0,376
5	4,974	0,769	0,419	10	9,921	0,490	0,945

Mejora en ubicación de la posición utilizando Downhill Simplex

La ubicación de los centros de las bandas utilizando la segunda derivada tiene el inconveniente que cuando el ancho de las bandas aumenta y el traslape es severo, la posición de los picos detectados se desvían del valor real.

Por esta razón se utiliza el valor obtenido de la desviación estándar mediante *LineSearch* y las posiciones calculadas con la segunda derivada para mejorar la aproximación de la posición de las bandas en el carril. Ejemplificando el trabajo del algoritmo se evalúa de nuevo para el carril de la figura 4.19.

En esta evaluación se eligió de nuevo el mismo carril utilizado anteriormente por el traslape severo existente en ese carril como caso especial. En la Figura 4.21 se muestra el error obtenido en píxeles para las bandas del carril, que como se muestra no excede el límite máximo considerado en este trabajo que es de 2 píxeles.

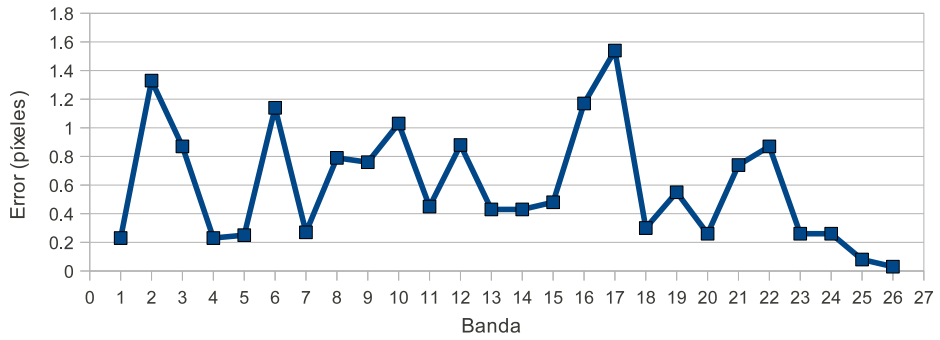


Figura 4.21: Error en la determinación de la posición utilizando Downhill Simplex.

Tabla 4.14: Posición aproximada de las bandas en un carril sintético con 26 bandas utilizando Downhill Simplex

Posición teórica	Posición obtenida	Posición teórica	Posición obtenida
17	16,77	403	402,57
70	68,67	416	415,52
78	78,87	423	424,17
116	115,67	440	441,54
140	140,25	485	485,30
154	154,14	587	587,55
182	181,73	642	641,84
228	227,21	753	752,26
237	237,76	762	762,87
271	269,97	835	834,74
279	279,94	862	862,26
321	321,45	902	902,08
327	327,88	977	977,03

Evaluación de carriles reales

El método de *LineSearch* necesita que las evaluaciones de la función objetivo tomen como base un carril sin distorsiones donde solamente se encuentren presentes las bandas, por lo que es necesario remover el fondo no deseado de los carriles reales con el algoritmo desarrollado en el proyecto anterior [7].

Antes de utilizar el ajuste de mínimos cuadrados se remueve el fondo de los carriles de acuerdo con la desviación estándar calculada mediante el espacio de escalas, por lo que antes de realizar el ajuste de la función objetivo se aplica un filtrado al carril real, para utilizarlo luego como entrada al programa.

La figura 4.22 presenta el proceso de ajuste mediante mínimos cuadrados. Para un carril real a) que presenta distorsión por fondo indeseado, cuya imagen de perfil es b) donde se aprecia las regiones que no son banda. A este perfil se le realiza la extracción del fondo con lo cual se obtiene el nuevo perfil filtrado c). Finalmente el ajuste de mínimos cuadrados se realiza con un vector simulado d) con bandas es las posiciones encontradas mediante la segunda derivada, de forma que se minimice la función objetivo con c) como referencia.

Puesto que el valor de la desviación estándar obtenida con el espacio de escalas puede estar alejada del valor real, se realiza un proceso iterativo utilizando el σ calculado en la iteración anterior del *LineSearch* para filtrar nuevamente el fondo de carril y aproximarlo por sumatoria de gaussianas hasta que se estabilice el valor de la desviación estándar obtenida.

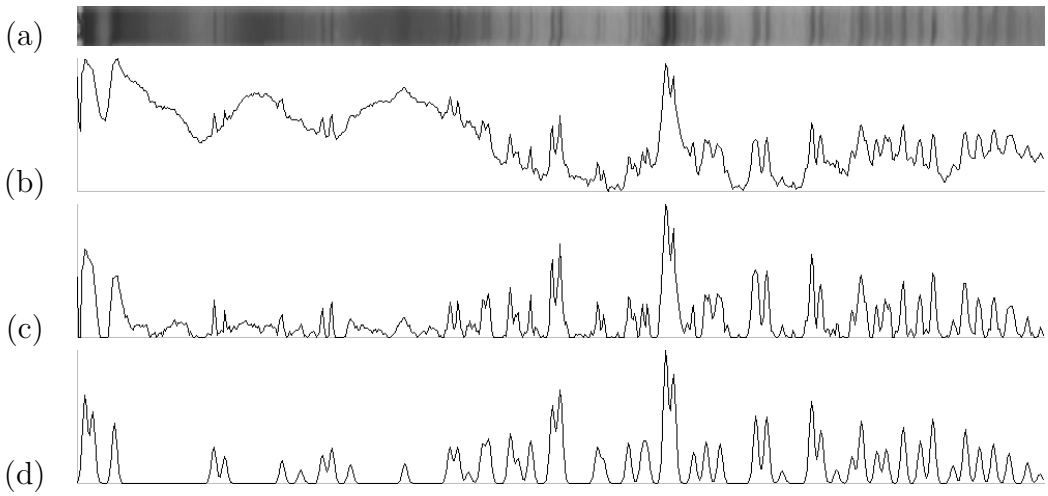


Figura 4.22: Proceso de ajuste del carril mediante mínimos cuadrados. a) Carril real, b) perfil, c) perfil sin fondo, d) perfil simulado de mejor ajuste.

Así el algoritmo propuesto en este trabajo tiene como resultado para el carril de la figura 4.22 la detección de 50 bandas y una desviación estándar de 1.813 para el mejor ajuste.

Más detalles sobre la evaluación de la detección de bandas se presentan en la tesis [22].

4.5 Implementación de la DGGE para el análisis de la diversidad genética bacteriana en muestras ambientales

A continuación se muestran los resultados obtenidos en la implementación de una estrategia para el estudio molecular de comunidades microbianas de muestras ambientales. La estrategia incluye la toma y manejo de la muestra, la extracción del ADN total, la amplificación del gen ARNr 16S por PCR y la resolución del ADN amplificado por Electroforesis en Gel de Gradiente Desnaturalizante.

4.5.1 Extracción de ADN

Las figuras 4.23 y 4.24 ilustran los resultados obtenidos de la extracción de ADN total de suelo y muestras de agua. Como se ve en ambas figuras, y como lo confirman los tratamientos posteriores, el ADN total obtenido es de buena calidad y esta en suficiente cantidad para estudios ulteriores.

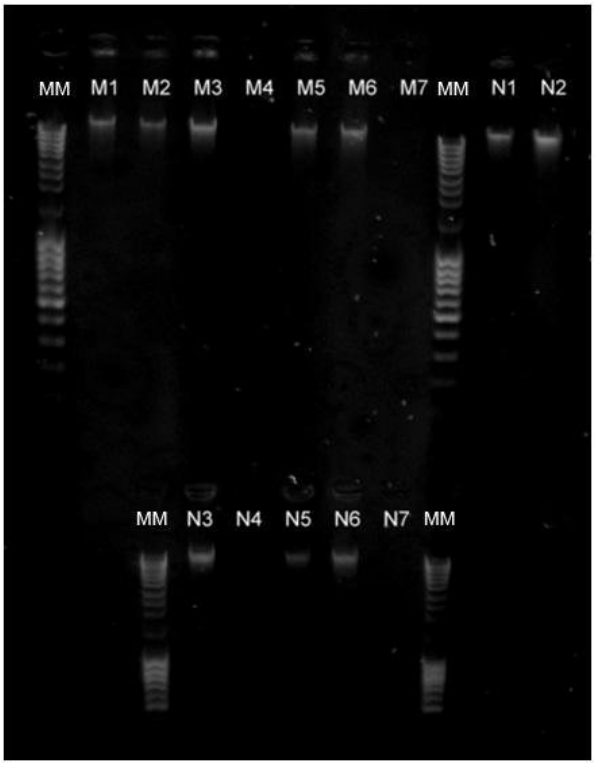


Figura 4.23: Extracciones de ADN total de muestras de suelos (de M1 a M7 y de N1 a N7, MM: marcador molecular.

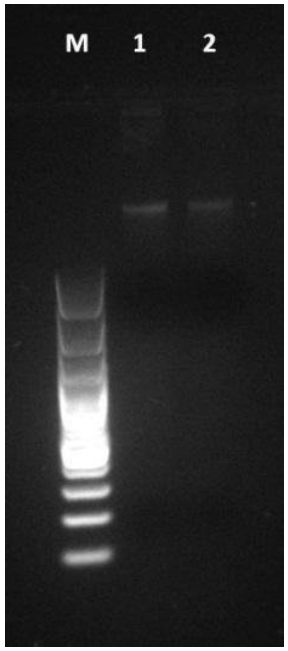


Figura 4.24: Extracciones de ADN total de muestras de agua (M: marcador molecular).

4.5.2 PCR

Las figuras 4.25 y 4.26 muestran el resultado de las valoraciones de los ADN extraídos de agua y suelo respectivamente mediante la técnica de PCR. El tamaño del fragmento de ADN que se observa en las distintas muestras corresponden al tamaño del fragmento esperado que es de aproximadamente 233pb (figura 4.25).

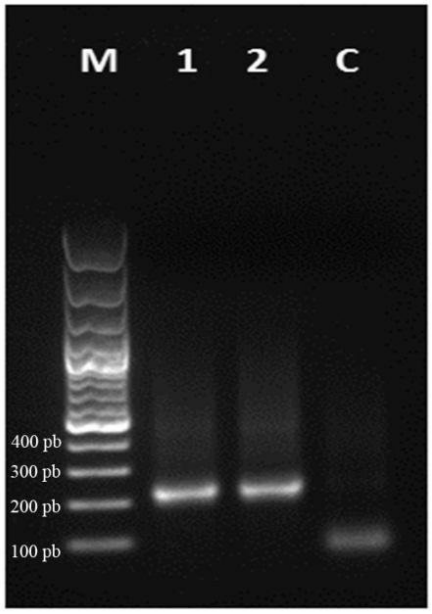


Figura 4.25: Amplificación por PCR del gen 16S del ARNr de eubacterias a partir de ADN total extraído de muestras de agua (M: marcador molecular, C: control).

4.5.3 DGGE

Las figuras 4.27 y 4.28 ilustran los resultados obtenidos en los geles de gradiente desnaturalizante con muestras de diferentes suelos (figura 4.27) y agua (figura 4.28). Las distintas bandas observadas en los diferentes carriles sugieren en buena medida la diversidad bacteriana de lugar y el grosor o la intensidad de la bandala abundancia de esa posible especie.

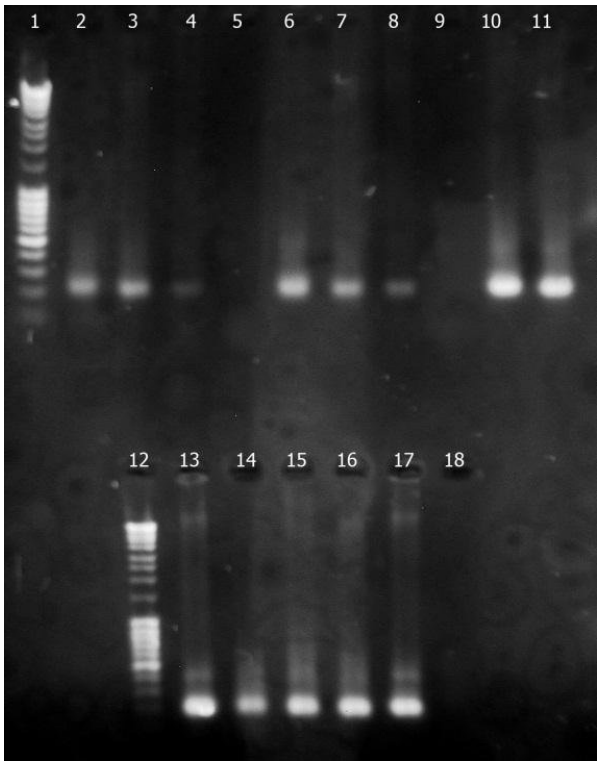


Figura 4.26: Amplificación por PCR del gen 16S del ARNr de eubacterias a partir de ADN total extraído de muestras de suelo (carril 1 y 12: marcador molecular; carriles 5, 9 y 18 corresponden a controles).

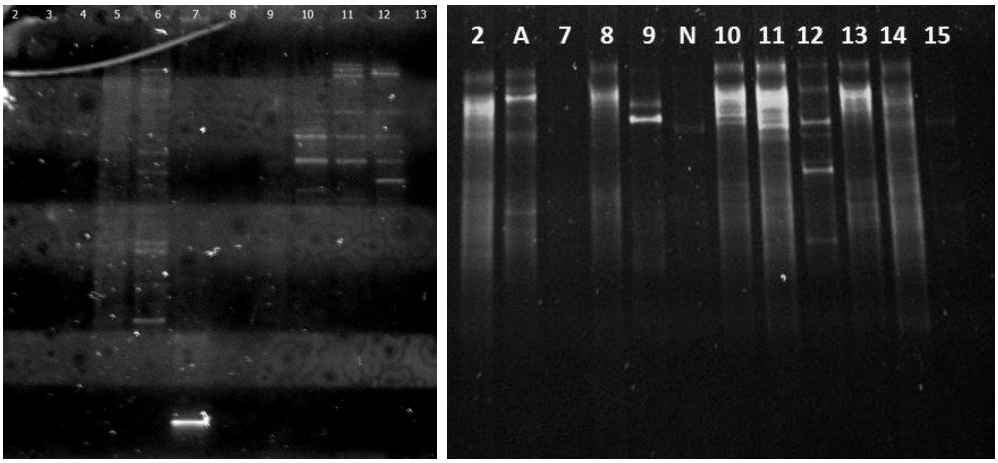


Figura 4.27: Geles de gradiente desnaturalizante de productos amplificados del gen 16s bacterianos procedentes de muestras de distintos suelos.

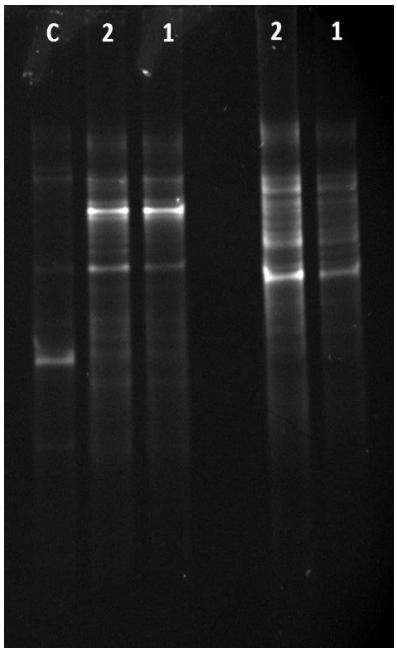


Figura 4.28: Geles de gradiente desnaturalizante de productos amplificados del gen 16s bacterianos procedentes de muestras de agua.

Capítulo 5

Conclusiones y recomendaciones

El objetivo general del presente proyecto ha sido incorporar a la herramienta desarrollada en los proyectos antecesores opciones avanzadas para el manejo de meta-información y para el análisis de imágenes de geles de electroforesis utilizadas en la caracterización molecular de organismos, de modo que se mejore la usabilidad del sistema en laboratorios de biología molecular.

Se incorporaron al sistema existente más funciones para el ingreso de usuarios y mantenimiento de datos. Todo sistema debe contar con políticas de manejo de seguridad, por lo que es recomendable en una version posterior plantear una política de manejo de claves seguras, que expiren, no puedan ser reutilizadas, y no contenga palabras fácilmente decifrables. El manejo de datos encriptados es otra majora que se debe dar a este proyecto, esto incrementa la seguridad a todo nivel, de forma que los datos guardados no sean vulnerables.

Se propuso el diseñó del sistema de bases de datos distribuidas que permite realizar búsquedas de geles y carriles en las bases de datos de los nodos participantes permitiendo realizar tareas de minería de datos.

En cuanto a la inclusión de modos adicionales para la manipulación de imágenes, interacción con el sistema y presentación de la información, se propuso el diseño de un módulo para minería de datos con capacidades de clasificación y clustering. Esto permitirá eventualmente identificar automáticamente los carriles de control con base en datos almacenados.

Se presentó el diseño de un sistema de adquisición mediante el ajuste multiparamétrico de la captura, donde se demostró el incremento de las medidas de contraste y reducción del nivel de ruido estimado en las imágenes capturadas mediante la optimización del sistema. El análisis estadístico realizado para diseñar dicho sistema le permite actuar y tomar decisiones iniciando en un punto concreto con la evaluación de una imagen de entrada generada con base en el rango solicitado, y en menos de 3 segundos, sin más iteraciones que las diez requeridas por el sistema para realizar comparaciones de calidad y seleccionar la mejor de las diez imágenes generadas, llegar al resultado.

De los algoritmos de fusión implementados, el algoritmo de fusión de exposición obtuvo un mayor incremento de las medidas de calidad en las imágenes que el algoritmo de fusión simple. Se demostró una ampliación en el rango dinámico y un incremento en las medidas de contraste de las imágenes de salida con respecto a las imágenes de entrada. Mediante el sistema global implementado se logró incrementar la nitidez de las imágenes de salida y objetivamente se obtuvieron mayores índices de calidad en las imágenes capturadas por el sistema que las obtenidas de manera manual.

Para el objetivo asociado a la rectificación y normalización de las imágenes de geles se siguieron dos caminos: la detección de carriles y la difusión coherente para detectar el efecto sonrisa.

Se ha propuesto utilizar un Modelo Activo de Forma capaz de ajustarse a los bordes de los carriles en una imagen de un gel de electroforesis, de forma que represente las distorsiones ópticas que éstos puedan presentar. Para esta construcción se parte de modelos de distorsión óptica teóricos, que han podido pueden ser modelados utilizando tan solo 2 dimensiones del espacio proyectado aún cuando el espacio original es de 220 dimensiones (110 puntos). Esta reducción implica una menor cantidad de información que debe ser procesada traduciéndose en menor cantidad de cálculos y de datos almacenados.

Para ajustar las formas a los bordes de los carriles se propone utilizar la propuesta basada en el cálculo del gradiente. Con el fin de ubicar las formas sobre la imagen es necesario encontrar el ancho de los carriles. El método de la autocorrelación de las columnas del gradiente resulta preciso (diferencias de un píxel) hasta con un 15% de ruido blanco gaussiano inducido en la imagen, aunque queda espacio de investigación en métodos para seleccionar los máximos locales correctos.

Se logra comprobar que el proceso iterativo de ajuste de las formas a la imagen es necesario para reducir la desviación estándar que presentan los carriles luego del proceso de rectificación. En cuanto a esta reducción se observó que los carriles con mayor desviación estándar previa a la rectificación (aproximadamente 8 píxeles) tienen una mayor reducción (de 6-7 unidades) que aquellos con baja desviación inicial (1-2 píxeles).

El método propuesto para reducción del efecto sonrisa empleando difusión mejorada en coherencia con filtrado de fase permitió corregir la distorsión del efecto sonrisa para una imagen en más del 99%, sin embargo, hay espacio para mejoras puesto que el sistema obtenido es sensible a los defectos en la imágenes (rayones, deformaciones y puntos) así como a las anotaciones en las mismas.

Los algoritmos requeridos para eliminar esas características requieren estrategias de reconocimiento de patrones que escapan al marco de este proyecto. Además, a pesar de que para una imagen particular es posible encontrar los parámetros óptimos, existe una fuerte relación entre los parámetros utilizados para el algoritmo y la calidad de la detección. Más investigación es requerida para intentar deducir los parámetros del algoritmo a partir de características de la imagen.

Para el análisis de los geles se incorporó la información de modelos de bandas, propia

del contexto de cromatografía y electroforesis. Se propuso una estrategia para realizar la ubicación automática de bandas de los geles de electroforesis. Se comprobó que es posible modelar la distribución de intensidad que caracteriza el perfil de una banda mediante una función gaussiana, para la cual su valor medio determina la ubicación central de la banda a lo largo del carril. Esto permitió modelar la distribución de bandas como una sumatoria de funciones gaussianas para así encontrar la ubicación de las mismas mediante una estrategia de optimización de parámetros y una función objetivo. Para encontrar la ubicación de las bandas es necesario realizar la extracción del fondo de la imagen.

Por otra parte, es posible encontrar los parámetros de las bandas estableciendo como función objetivo el error cuadrático medio entre la sumatoria de funciones gaussianas y la intensidad promedio del carril. Se comprobó que para encontrar los parámetros que minimizan esta función objetivo es necesario utilizar un optimizador basado únicamente en la información dada por la función y una estrategia de generación de puntos multidimensionales iniciales a optimizar, lo cual es posible combinando los principios de los algoritmos genéticos PESA y los frentes de Pareto con una estrategia de optimización como Downhill Simplex.

Para encontrar el ancho de las bandas requerido para la optimización se utiliza un análisis del carril en el espacio de escalas. Si bien se procuró reutilizar dicho análisis en el posicionamiento de las bandas, se demostró matemáticamente que no es posible por la influencia de los lóbulos laterales de la segunda derivada, por lo que es necesario utilizar una estrategia de aproximación para ubicarlas en los puntos correspondientes a los máximos, lo cual se logró al implementar un optimización multidimensional de los datos iniciales.

El ajuste que utiliza minimización lineal en conjunto con la segunda derivada para posición de las bandas permite aproximar mejor el valor de la desviación estándar, ya que permite tomar en cuenta las bandas del carril que no serían distinguibles por el traslape entre ellas. Se comprobó que cuando pueden detectarse las bandas presentes en el carril la desviación estándar será aproximada correctamente con error menor a 1 píxel permitido aún cuando existe traslape.

Para obtener un sistema automatizado es necesario utilizar un método que brinde una aproximación inicial de la desviación estándar, como es en este caso el espacio de escalas, porque se necesita un valor cercano al real para realizar el filtrado inicial antes de realizar las aproximaciones de la función objetivo con el carril filtrado, y así corregir el valor de la desviación estándar.

La detección de la posición inicial de las bandas permite disminuir considerablemente el tiempo necesario para la caracterización de las bandas en comparación a un método puro de optimización, en especial cuando las ventanas del carril analizadas contienen varias bandas.

Finalmente, se logró establecer el protocolo para generación de electroforesis en gel de gradiente desnaturalizante. Sus características hacen que el sistema propuesto los pueda integrar de forma natural.

Capítulo 6

Aportes

Se ha propuesto un subsistema de captura que mejora el rango dinámico de las imágenes por medio de la fusión de imágenes capturadas con diferentes parámetros de la cámara, reduciendo el ruido final y maximizando el contraste de las imánenes. Dentro de esta propuesta se incluye un método de estimación del ruido novedoso.

Mejoras considerables en el algoritmo de detección de carriles y rectificación de las imágenes fueron alcanzados utilizando modelos de forma y métodos de autocorrelación para la determinación automática del ancho del carril.

Para la detección del efecto sonrisa se realizó una modificación al algoritmo de difusión mejorada en coherencia para incluir un filtrado de orientación, que permite obligar al proceso de difusión a seguir el patrón de las bandas de los geles por encima del patrón de los carriles, que son visualmente dominantes y los únicos detectables con los algorimos existentes previamente.

La detección de bandas es un proceso central en el análisis de imágenes de electroforesis y en este proyecto se aporta una metodología híbrida de métodos de optimización para encontrar la posición y amplitud de las bandas y métodos basados en el análisis de espacio de escalas para encontrar el ancho de dichas bandas.

Se diseñó el sistema de bases de datos distribuidas que brinda acceso a los datos recolectados, lo que incluye el esquema de exportación, el proceso de consultas, un sistema de caché para reducir el intercambio de datos y el protocolo para incorporar nuevos nodos al sistema distribuido.

Se diseñó además un módulo de minería de datos para proveer acceso a los datos usando técnicas de clasificación por árboles de decisión y agrupamiento.

Finalmente, se logró establecer el protocolo para producir electroforesis en gel de gradiente desnaturalizante, que se utilizó en el análisis de diversidad genética bacteriana en muestras ambientales (suelo y agua).

Bibliografía

[1] Antonio Aguilar Bravo. Detección y corrección del efecto sonrisa en imágenes de geles de electroforesis utilizando modelos activos de forma acoplados. Tesis de licenciatura, Escuela de Ingeniería Electrónica, Instituto Tecnológico de Costa Rica, Cartago, Abril 2010.

[2] F. Xabier Albizuri. *Procesamiento de imagen digital*. Universidad Euskal Herriko, España, Enero 2010.

[3] D. F. Allen, J. S. Bashkin, Hong Guo, S. E. Shadle, W. K. Pogozelski, y T. D. Tullius. Quantitative analysis of electrophoresis data: novel curve fitting methodology and its application to the determination of a protein–DNA binding constant. *Nucleic Acids Research*, 25(4), 1997.

[4] Pedro Alpízar Salas y Pablo Alvarado-Moya. Anisotropic Diffusion on Electrophoresis Images. In *Proceedings of the Conference on Technologies for Sustainable Development TSD2011*, Cartago, Costa Rica, 2011. URL http://www.ie.itcr.ac.cr/palvarado/papers/TSD2011_Paper_06.pdf.

[5] Pedro Elías Alpízar Salas. Optimización de la corrección del efecto sonrisa en imágenes de geles de electroforesis. Tesis de licenciatura, Escuela de Ingeniería Electrónica, Instituto Tecnológico de Costa Rica, Cartago, Diciembre 2010.

[6] P. Alvarado y A. Salazar. Análisis automatizado de patrones de ADN para la caracerización molecular. Informe final. Actividad de fortalecimiento de la investigación, Instituto Tecnológico de Costa Rica, Marzo 2008.

[7] P. Alvarado, A. Salazar, O. Murillo, F. Rojas, y J. Peraza. Análisis por computador de imágenes de geles de electroforesis para la caracterización molecular de organismos. Informe final, Instituto Tecnológico de Costa Rica, Abril 2010.

[8] Pablo Alvarado-Moya. *Segmentation of color images for interactive 3D object retrieval*. PhD thesis, RWTH-Aachen, Alemania, 2004.

[9] Bryant Esteban Álvarez Canales. Mejoramiento de contraste y razón señal a ruido de imágenes digitales de geles de electroforesis por medio de fusión y ajuste multiparamétrico de la captura. Tesis de licenciatura, Escuela de Ingeniería Electrónica, Instituto Tecnológico de Costa Rica, Cartago, Noviembre 2011.

[10] R.D. Appel, A. Bairoch, J.C. Sanchez, J.R. Vargas, O. Golaz, C. Pasquali, y D.F. Hochstrasser. Federated 2-de database: a simple means of publishing 2-de data. *Electrophoresis*, 17:540–546, 1996. URL http://world-2dpage.expasy.org/swiss-2dpage/docs/fed-rules.html.

[11] D.G. Bailey y B.C. Christie. Processing of dna and protein electrophoresis gels by image analysis. In *Proceedings of the second New Zealand Conference on Image and Vision Computing*, págs. 2.2.1–2.2.8, Palmerston North, August 1994.

[12] I. Bajla, I. Holländer, y K. Burg. Improvement of electrophoretic gel image analysis. *Measurement Science Review, Slovak Academy of Science*, 1(1), 2001.

[13] Pablo Barrantes-Chaves. Detección de carriles y rectificación de imágenes de geles de electroforesis utilizando modelos activos de forma. Tesis de licenciatura, Escuela de Ingeniería Electrónica, Instituto Tecnológico de Costa Rica, Cartago, Junio 2010.

[14] Pablo Barrantes-Chaves y Pablo Alvarado-Moya. Lane Detection on Gel Electrophoresis Images using Active Shape Models. In *Proceedings of the Conference on Technologies for Sustainable Development TSD2011*, Cartago, Costa Rica, 2011. URL http://www.ie.itcr.ac.cr/palvarado/papers/TSD2011_Paper_03.pdf.

[15] J. M. Berg, J. L. Tymoczko, y L. Stryer. *Biochemistry*. W. H. Freeman and Company, 5th edición, 2002. URL http://www.ncbi.nlm.nih.gov/books/NBK21154/.

[16] J.F. Canny. A computational approach to edge detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 8(6):679–698, August 1986.

[17] N. Carballido, A. D. García Hernández, V. L. Ballarían, y J. I. Pastore. Desarrollo de técnicas de procesamiento digital para la optimización de imágenes de fragmentos de ADN obtenidas en estudios de identificación humana. Technical report, Facultad de Ingeniería.Universidad Nacional de Mar de Plata, Argentina, 2005.

[18] T.˜F. Cootes, C.˜J. Taylor, D.˜H. Cooper, y J. Graham. Active Shape Models — Their Training and Application. *Computer Vision and Image Understanding*, 61(1):38–59, January 1995.

[19] David W Corne, Joshua D Knowles, y Martin J Oates. The Pareto Envelope-based Selection Algorithm for Multiobjective Optimization. *Decision Analysis*, 1917(Mcdm):839–848, 2000. URL http://www.springerlink.com/index/q576765808l68p34.pdf.

[20] Peter Dörfler y Pablo Alvarado-Moya. LTI-Lib - A C++ Open Source Computer Vision Library. In Karl-Friedrich Kraiss, editor, *Advanced Man-Machine Interaction. Fundamentals and Implementation*, chapter LTI-Lib -, págs. 399–421. Springer-Verlag, 2006. URL http://www.springer.com/sgw/cda/frontpage/0,11855,4-154-22-112579645-detailsPage%253Dppmmedia%257CaboutThisBook%257CaboutThisBook,00.html.

[21] C. G. Enke y T. A. Nieman. Signal to noise ratio enhancement by least-squares polynomial smoothing. *Analytical Chemistry*, 48(8):705A–712A, July 1976.

[22] Randall José Esquivel Alvarado. Optimización de la detección de cantidad, ancho y posición de bandas en imágenes de geles de electroforesis. Tesis de licenciatura, Escuela de Ingeniería Electrónica, Instituto Tecnológico de Costa Rica, Cartago, Junio 2011.

[23] Mark Everingham, Henk Muller, y Barry Thomas. Evaluating image segmentation algorithms using the pareto front. In *Proc. Seventh European Conf. Computer Vision*, págs. 34–48, 2002.

[24] A Farnleitner, T Hein, G Kavka, R. Mach, y C. Winter. Longitudinal changes in the bacterial community composition of the danube river: a whole-river approach. *Applied And Environmental Microbiology*, 73(2):421–431, 2007.

[25] A. Felinger. *Data Analysis and Signal Processing in Chromatography*. Elsevier, 1998.

[26] Edison Fernández Alvarado. Estimación de la desviación estándar para un modelo gaussiano del perfil de bandas en imágenes de geles de electroforesis utilizando información de múltiples escalas. Tesis de licenciatura, Escuela de Ingeniería Electrónica, Instituto Tecnológico de Costa Rica, Cartago, Noviembre 2010.

[27] C.L.W.T. Freeman y R.S.S.B. Kang. Noise estimation from a single image. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2006.

[28] Ronald García. Corrección de distorsión geométrica y detección de carriles en imágenes de geles de electroforesis para la caracterización molecular de organismos por computador. Tesis de licenciatura, Escuela de Ingeniería Electrónica, Instituto Tecnológico de Costa Rica, Cartago, Junio 2009.

[29] Chris Glasbey, Leila Vali, y John Gustafsson. A statistical model for unwarping of 1-D electrophoresis gels. *Electrophoresis*, 26(22):415–419, November 2005.

[30] R.C. González y R.E. Woods. *Digital Image Processing*. Prentice-Hall, 3rd edición, 2008.

[31] A. Goshtasby y W. D. O'Neill. Curve fitting by a sum of gaussians. *Graphical Models and Image Processing*, 54(4):281–288, July 1994. URL http://www.engineering.wright.edu/~agoshtas/GMIP94.pdf.

[32] A. Hornberg, editor. *Handbook of Machine Vision*. Wiley-VCH, 2008.

[33] H. Irshad, M. Kamran, A.B. Siddiqui, y A. Hussain. Image fusion using computational intelligence: A survey. In *2009 Second International Conference on Environmental and Computer Science*, págs. 128–132. IEEE, 2009.

[34] P. Jung-Me, C.G. Looney, y C. Hui-Chuan. Fast connected component labeling algorithm using a divide and conquer technique. In *CATA 2000 Conference on Computers and Their Applications*, págs. 373–376, 2000.

[35] B. Jähne. *Digital Image Processing*. Springer-Verlag, 6th revised and extended edition edición, 2005.

[36] KDNuggets. Data mining software suites [online]. 2012 [visitado el 12 de septiembre de 2012]. URL http://www.kdnuggets.com/software/suites.html.

[37] Lei Li y Terence P Speed. Parametric Deconvolution of Positive Spike Trains. *Annals of Statistics*, 28(5):1279–1301, 2000. URL http://www.jstor.org/stable/2674094.

[38] TotalLab Limited. Phoretix 1d pro – overview [online]. 2009 [visitado el 12 de septiembre de 2012]. URL http://www.totallab.com/products/1dpro.

[39] T Lindeberg. Scale-space theory: A basic tool for analysing structures at different scales. *Journal of Applied Statistics*, 21(2):224–270, 1994.

[40] T Lindeberg. Feature Detection with Automatic Scale Selection. *International Journal of Computer Vision*, 30(2):79–116, November 1998.

[41] Tony Lindeberg. *Discrete Scale-Space Theory and the Scale-Space Primal Sketch*. PhD thesis, Royal Institute of Technology, 1991. URL http://ftp.nada.kth.se/pub/CVAP/reports/Lin91-PhDthesis-missing-figures.pdf.

[42] Alex Logan, Galina Havin, y Larry Arend. Luminance contrast [online]. 2011 [visitado el 30 de noviembre de 2011]. URL http://colorusage.arc.nasa.gov/luminance_cont.php.

[43] H. Lu, H. Zhang, S. Yang, y Z. Zheng. Camera parameters auto-adjusting technique for robust robot vision. In *Robotics and Automation (ICRA), 2010 IEEE International Conference on*, págs. 1518–1523. IEEE, 2010.

[44] John A Luckey, Tracy B Norris, y Lloyd M Smith. Analysis of resolution in DNA sequencing by capillary gel electrophoresis. *Journal of Physical Chemistry*, 97(12):3067–3075, 1993. URL http://pubs.acs.org/doi/abs/10.1021/j100114a038.

[45] Hall M., E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, y I. H. Witten. The WEKA Data Mining Software: An Update. *SIGKDD Explorations*, 11(1), 2009.

[46] Matrix Science Ltd. Mascot integra [online]. 2012 [visitado el 12 de septiembre de 2012]. URL http://www.matrixscience.com/integra.html.

[47] T. Mertens, J. Kautz, y F. van Reeth. Exposure fusion. In *Computer Graphics and Applications, 2007. PG'07. 15th Pacific Conference on*, págs. 382–390. IEEE, 2007.

[48] G. Muyzer, E.C. De Waal, y A.G. Uitterlinden. Profiling of complex microbial populations by denaturing gradient gel electrophoresis analysis of polymerase chain reaction-amplified genes coding for 16S rRNA. *Applied and Environmental Microbiology*, 3(59):695–700, 1993.

[49] A. O. Olinaran, W.H.L. Stafford, D.A. Cowan, D. Pillay, y B. Pillay. Microbial community profiling in cis- and trans-dichloroethene enrichment systems using denaturing gradient gel electrophoresis. *Journal of Microbiology and Biotechnology*, 17(4):560–570, 2007.

[50] Oracle corporation. Mysql 5.5 reference manual – c api support for multiple statement execution [online]. 2012 [visitado el 12 de septiembre de 2012]. URL http://dev.mysql.com/doc/refman/5.5/en/c-api-multiple-queries.html.

[51] E. Peli. Contrast in complex images. *JOSA A*, 7(10):2032–2040, 1990.

[52] G. Piella. Image fusion for enhanced visualization: a variational approach. *International journal of computer vision*, 83(1):1–11, 2009.

[53] William H. Press, Saul A. Teukolsky, William T. Vetterling, y Brian P. Flannery. *Numerical Recipes. The Art of Scientific Computing*. Cambridge University Press, third edición, 2007.

[54] J. G. Proakis y D. G. Manolakis. *Tratamiento Digital de Señales*. Prentice Hall, 1998.

[55] J. R. Quinlan. *C4.5: Programs for Machine Learning*. Morgan Kaufmann Publishers, 1993.

[56] Rahimi S. y Haug F. *Distributed Management Systems: A Practical Approach*. John Wiley & Sons, 2010.

[57] M. Sonka, V. Hlavac, y R. Boyle. *Image processing, analysis and machine vision*. PWS Publishing, 2nd edición, 1999.

[58] David Soto-Vásquez y Pablo Alvarado-Moya. Automatic detection of bands in electrophoresis gel images by means of optimization of a target function. In *Proceedings of the Conference on Technologies for Sustainable Development TSD2011*, Cartago, Costa Rica, 2011. URL http://www.ie.itcr.ac.cr/palvarado/papers/TSD2011_Paper_22.pdf.

[59] David Soto Vásquez. Detección automática de bandas en imágenes de geles de electroforesis por medio de la optimización de una función objetivo. Tesis de licenciatura, Escuela de Ingeniería Electrónica, Instituto Tecnológico de Costa Rica, Cartago, Junio 2010.

[60] The Apache Software Foundation. Apache tomcat 7.0 [online]. 2012 [visitado el 12 de septiembre de 2012]. URL http://tomcat.apache.org/.

[61] Q.K. Vuong, S.H. Yun, y S. Kim. A new auto exposure and auto white-balance algorithm to detect high dynamic range conditions using cmos technology. In *Proceedings of the World Congress on Engineering and Computer Science*. Citeseer, 2008.

[62] Joachim Weickert. *Anisotropic Diffusion in Image Processing*. B.G. Teubner Stuttgart, 1998.

[63] Joachim Weickert. Coherence-enhancing diffusion filtering. *International Journal of Computer Vision*, 31:111–127, 1999.

[64] Joachim Weickert y Joachim Scharr. A Scheme for Coherence-Enhancing Diffusion Filtering with Optimized Rotation Invariance. *Journal of Visual Communication and Image Representation*, 13, 2002.

[65] I.H. Witten y E. Frank. *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kauffmann Series in Data Management Systems. Elsevier, second edición, 2005.

Apéndice A

Artículos publicados y reconocimientos

Las referencias [4, 14, 58] corresponden a tres artículos presentados con avances de este proyecto al congreso *Technologies for Sustainable Development TSD2011*.

La tesis de Randall Esquivel [22], desarrollada en el contexto de este proyecto, fue galardonada con el Premio Asoelectrónica 2011 a los mejores Proyectos de Graduación de Ingeniería Electrónica.