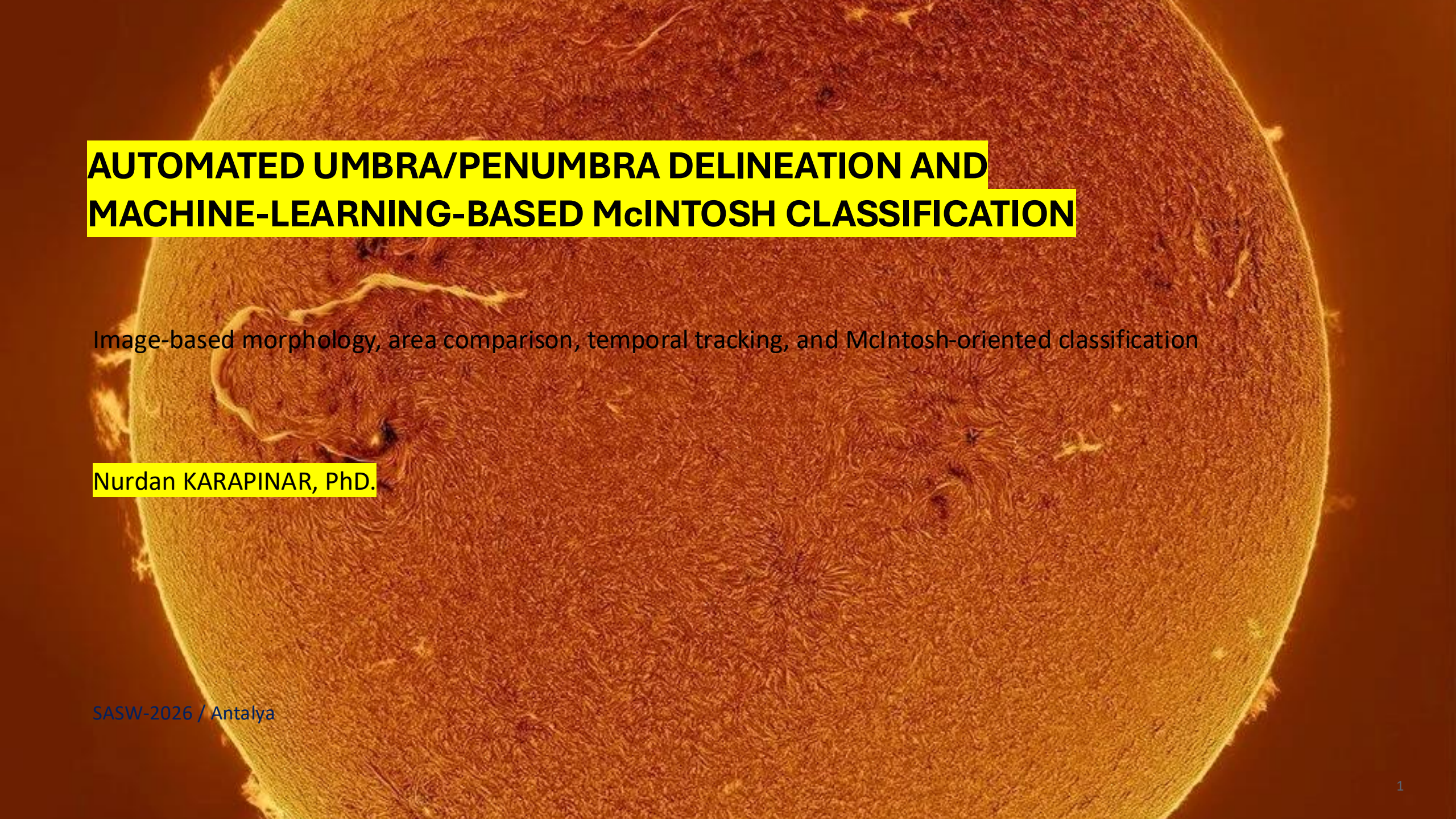


AUTOMATED UMBRA/PENUMBRA DELINEATION AND MACHINE-LEARNING-BASED McINTOSH CLASSIFICATION

Image-based morphology, area comparison, temporal tracking, and McIntosh-oriented classification

Nurdan KARAPINAR, PhD.

SASW-2026 / Antalya

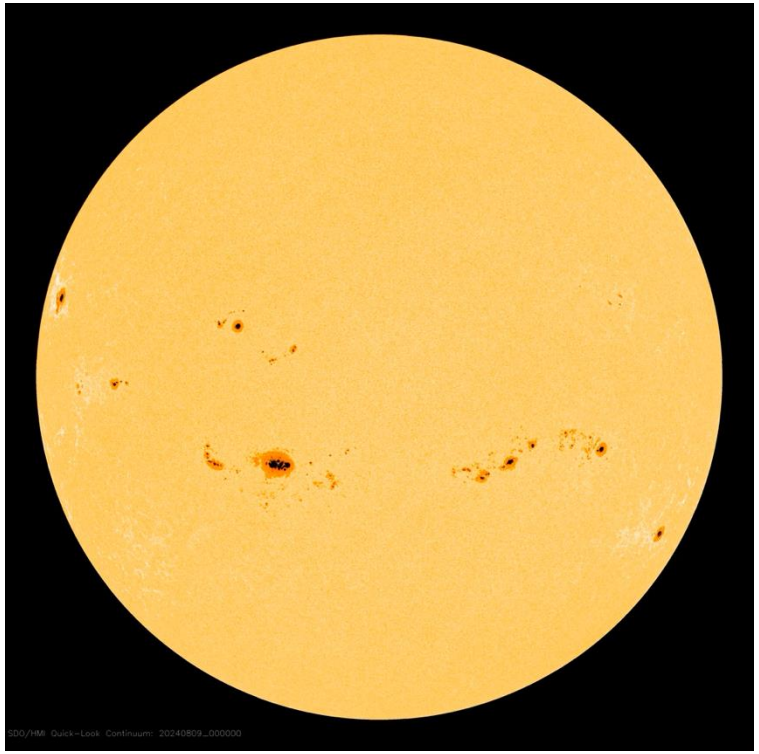
1. Scientific Objective

• Develop an automated, frame-by-frame pipeline for detecting sunspots area.

• To characterize total spot, umbra, and penumbra components.

• Quantify morphology using area, shape descriptors, and component fractions.

• Track regions through time and construct an image-based McIntosh-classification layer.



2024/08/09 - SDO, There is a usage example

2. Input Data and Temporal Sampling

Image geometry

1024 × 1024 pixels
Full-disk solar image

Temporal sampling

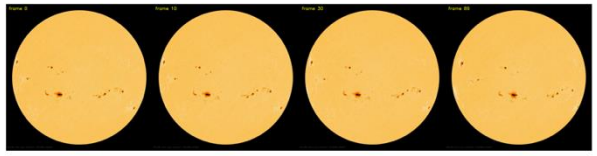
20 frames s^{-1}
$\Delta t = 0.05$ s

Sequence length

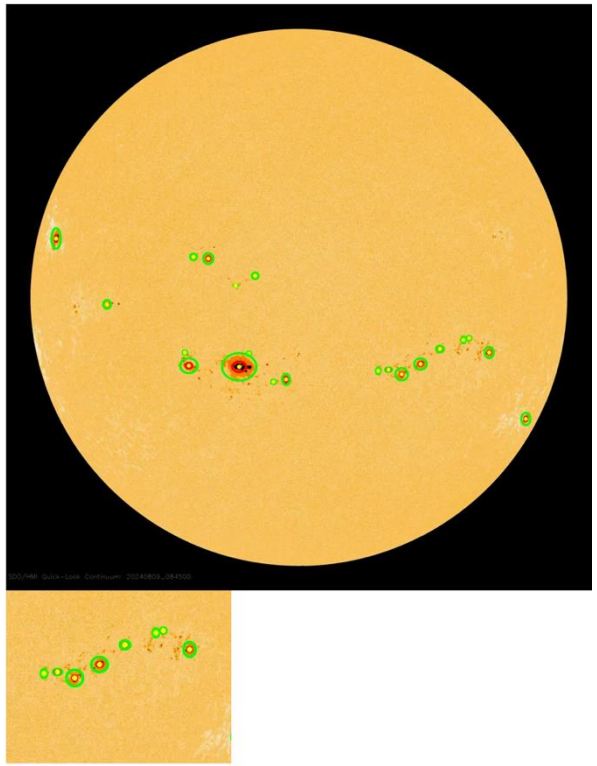
90 frames
4.50 s

Measurement rows

1,652 region-frame records







For instance mask

- The code reads every frame and preserves the native video cadence.
- The solar disk geometry is detected from the first frame and biased by default.
- The first-frame limb-darkening/quiet-Sun background is biased for subsequent frames to improve temporal consistency.
- This design is computationally efficient for a stable full-disk sequence, but assumes sufficiently stable imaging geometry.

3. Processing Pipeline

1 Frame ingestion

OpenCV video reader

2 Disk detection

Largest bright component

3 Limb-darkening correction

Radial median profile + Gaussian smoothing

4 Spot detection

Median–MAD thresholding + morphology

5 Umbra segmentation

Local Otsu threshold per region

6 Feature extraction

Area, shape, centroid, fractions

7 Temporal tracking

Centroid association

8 Outputs

MP4 + masks + Excel + diagnostics

4. Solar Disk and Limb-Darkening Correction

- The solar disk is identified as the largest bright connected component after thresholding, closing, hole filling, and small-object removal.
- The disk centre and radius are derived from the detected component.
- A radial median intensity profile is constructed over the disk.
- The profile is smoothed with a Gaussian kernel (RADIAL_SIGMA = 8).
- The normalized intensity image is obtained by dividing the observed image by the estimated quiet-Sun background.

Why this step matters

Limb darkening introduces a strong radial intensity gradient. Correcting it before thresholding makes a single robust intensity criterion more meaningful across disk position.

Current implementation

The first-frame background is reused by default. This reduces frame-to-frame threshold drift and computation, but requires stable pointing, scale, and illumination.

Key assumption

The radial profile represents the quiet-Sun intensity distribution adequately despite the presence of active regions.

5. Spot, Umbra, and Penumbra Segmentation

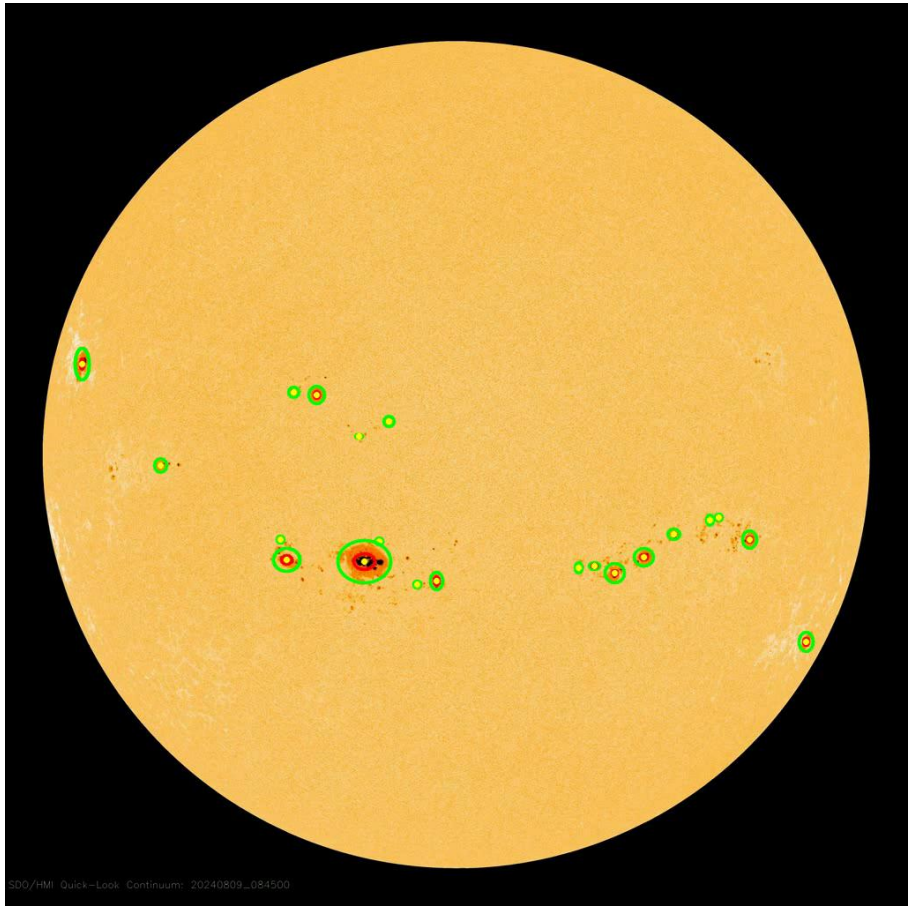
• Outer spot detection uses a robust median–MAD estimator on the limb-darkening-corrected image.

• Threshold: median − 2.5 × (1.4826 × MAD), constrained to the interval 0.70–0.995.

• Morphological hole filling, and small-object removal suppress isolated noise and fragmented detections.

• Each connected spot region is then segmented internally using a local Otsu threshold.

• The operational penumbra is defined as total detected spot area minus the detected umbral area.



Example visual output: green = outer boundary; red = umbral boundary; yellow = centroid.

Important terminology

"Penumbra" is an operational segmentation class in this pipeline, not a magnetically or spectroscopically calibrated penumbral measurement.

6. Morphological Measurements and Temporal Tracking

Scientific implication

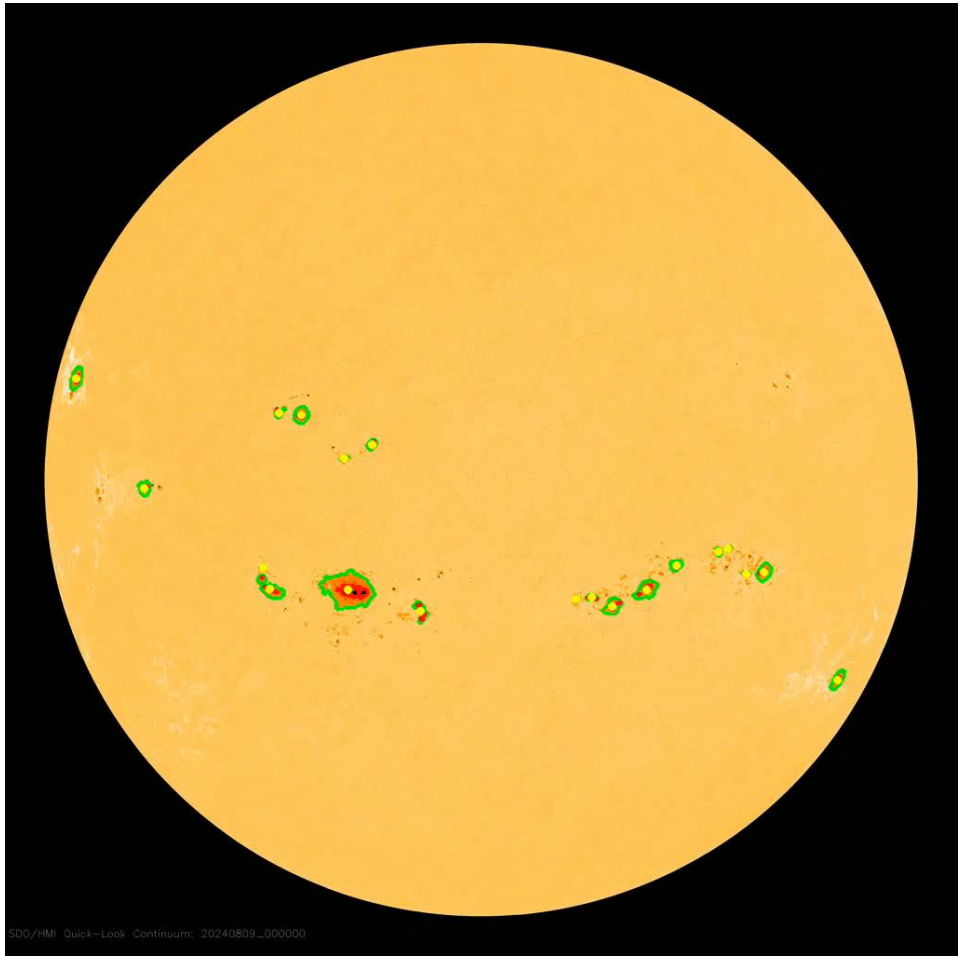
The tracker converts independent frame detections into time series for individual regions. This enables short-term analysis of area evolution and centroid motion.

However, centroid-only association does not explicitly model region splitting, merging, deformation, or differential motion. These cases can create track fragmentation or identity swaps.

Tracking model

Centroid-to-centroid association between consecutive frames.
Maximum association distance = 45 pixels.
Tracks may survive up to 5 missed frames.

McIntosh extension
Morphological features form the input vector for an optional Random Forest classifier when labelled training data are available.



7. What Is the Scientific Result?

- The pipeline converts full-disk intensity video into reproducible frame-resolved morphology and persistent region IDs.
- Total area is retained as an explicit comparative feature and converted to a projected MSH proxy.
- Umbra/penumbra partition, shape descriptors, and group extent provide a quantitative morphology vector.
- An image-based McIntosh Zpc proxy is now produced transparently; it is not an official NOAA/SWPC class.
- A supervised Random-Forest McIntosh classifier is available only when labelled training data are supplied.
- The current 4.5-s sequence demonstrates processing feasibility, not physical spot evolution timescales.

Central scientific statement

The extended pipeline establishes a quantitative morphology-to-classification framework. McIntosh classification must be calibrated and independently validated against expert or catalogue labels before being reported as a scientific ML classification result.

8. Scientific Interpretation and Publication Readiness

What can be claimed now

- Automated umbra/operational-penumbra delineation is implemented.
- Frame-resolved morphological measurements and temporal IDs are produced.
- Total area is explicitly available for region comparison.
- An image-based McIntosh proxy is generated transparently.
- The code contains an optional supervised Random-Forest stage.

What requires validation before publication

- Expert or catalogue McIntosh labels
- Independent train/test regions
- Calibration of proxy thresholds
- Foreshortening-corrected area
- Robust group definition and split/merge handling
- Longer time series
- Uncertainty estimates

The proposed title is therefore best supported as a classification framework until the supervised model is trained and validated.

9. Current Limitations and Required Development

Segmentation

- Threshold parameters are fixed globally.
- MAD thresholding is robust, but the choice of 2.5 MAD and the intensity clipping range is empirical.
- Local Otsu segmentation can be unstable for very small or low-contrast regions.
- Penumbra is inferred as a residual class.

Tracking

- Centroid distance is the only association criterion.
- No explicit split/merge handling.
- No motion prediction or shape similarity.
- Short-lived detections can fragment into multiple IDs.

Classification & physical interpretation

- The image-based McIntosh is a proxy because intensity data do not provide magnetic polarity.
- Empirical proxy thresholds require calibration against labelled McIntosh observations.
- Projected MSH is not foreshortening-corrected.
- No photometric calibration or uncertainty propagation.
- The 4.5-s sequence cannot establish lifetime, rotation, or long-term growth/decay.

10. Recommended Development Path

- Calibrate the image-based Zpc proxy against expert-labelled McIntosh groups.
- Build a region-level labelled dataset and avoid frame-level leakage between train/test sets.
- Train and evaluate the Random-Forest classifier using precision, recall, F1, confusion matrices, and group-wise validation.
- Add heliographic coordinates and foreshortening-corrected areas.
- Upgrade tracking with motion prediction plus explicit split/merge handling.
- Extend from seconds to days/weeks for emergence, growth, decay, recurrence, and rotation studies.
- Incorporate magnetograms when magnetic classification or polarity-dependent McIntosh validation is required.

Target end product

A validated automated framework linking sunspot segmentation, morphology, temporal tracking, and McIntosh classification with quantitatively documented uncertainty.

11. References

1. McIntosh, P. S. (1990). The classification of sunspot groups. Solar Physics, 125, 251–267. DOI: 10.1007/BF00158405.

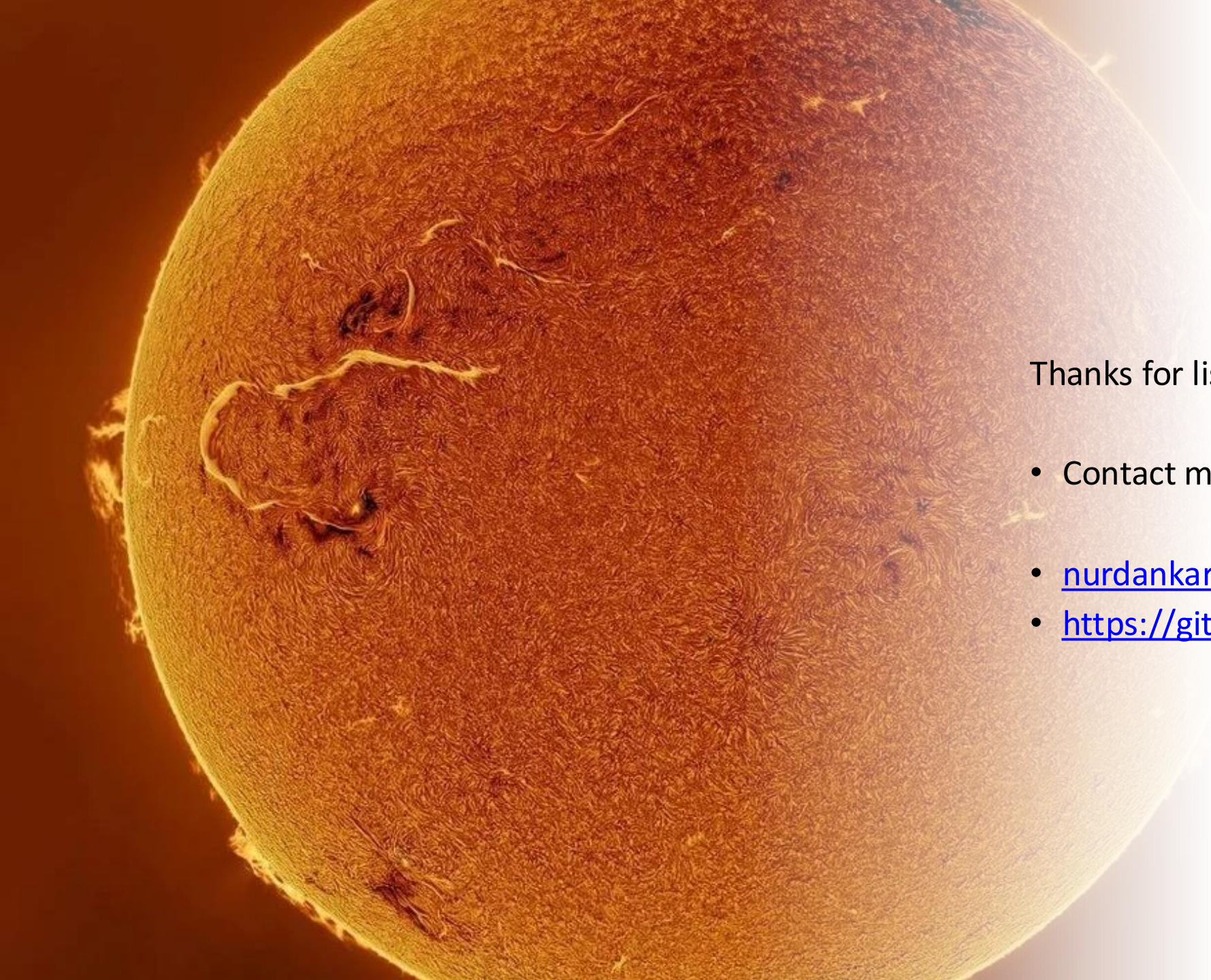
2. Colak, T., & Qahwaji, R. (2008). Automated McIntosh-based classification of sunspot groups using MDI images. Solar Physics, 248, 277–296. DOI: 10.1007/s11207-007-9094-3.

3. Sayez, M., et al. (2023). SunSCC: Segmenting, Grouping and Classifying Sunspots From Ground-Based Observations Using Deep Learning. Journal of Geophysical Research: Space Physics. DOI: 10.1029/2023JA031548.

4. Deng, X., et al. (2025). Deep learning-based McIntosh classification of sunspot groups. Astronomy and Computing, 53, 100995. DOI: 10.1016/j.ascom.2025.100995.

5. Colak, T., & Qahwaji, R. (2009). Automated solar activity prediction: A hybrid computer platform using machine learning and solar imaging for automated prediction of solar flares. Space Weather, 7, S06001. DOI: 10.1029/2008SW000401.

6. NOAA / SWPC educational material on sunspot classification and McIntosh terminology.



Thanks for listening ☺

- Contact me;

- nurdankarapinar@gmail.com
- https://github.com/nurdankarapinar

McIntosh Classification and the New Classification Layer

McIntosh Zpc

Z — modified Zürich evolutionary/group class

p — penumbra characteristics of the largest spot

c — compactness / distribution of spots within the group

The standard system combines these three components into admissible sunspot-group classes.

Features extracted by this code

Total area and projected MSH
Umbra and penumbra areas
Umbra / penumbra fractions
U/P ratio
Width, height, aspect ratio
Equivalent diameter and perimeter
Circularity and group extent
Fragmentation index

Scientific caveat

The current intensity-only image does not directly provide magnetic polarity.

Therefore the code reports an Estimated_McIntosh_Class as an image-based proxy.

It must not be presented as an official NOAA/SWPC class until calibrated and independently validated.

McIntosh scheme: McIntosh (1990), Solar Physics, 125, 251–267. DOI: 10.1007/BF00158405

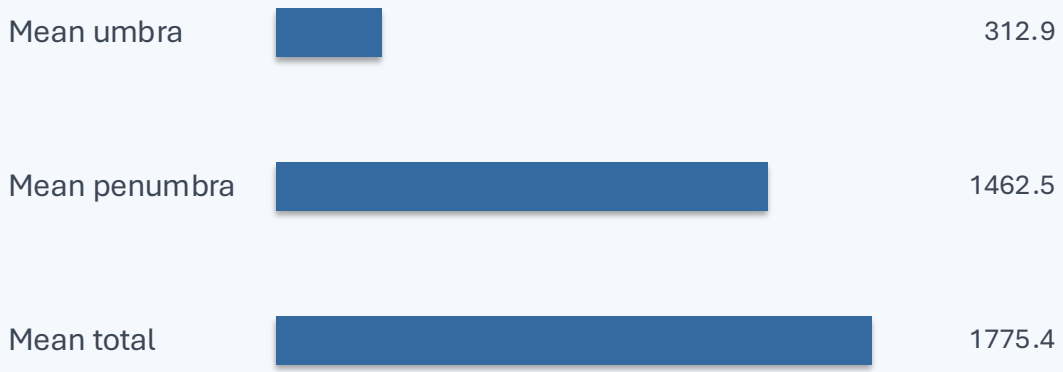
Area-Based Comparison and Current Sequence Results

Measured sequence

1,652 region-frame measurement rows
40 Region_ID tracks
13 tracks persist through all 90 frames
18.36 detected regions per frame on average
Total detected spot area per frame: 4,249–5,827 px^2
Mean umbra fraction: 0.199

R12 — largest tracked region
Mean total area: 1,775.4 px^2
Maximum total area: 1,912 px^2
Mean umbra area: 312.9 px^2
Mean penumbra area: 1,462.5 px^2

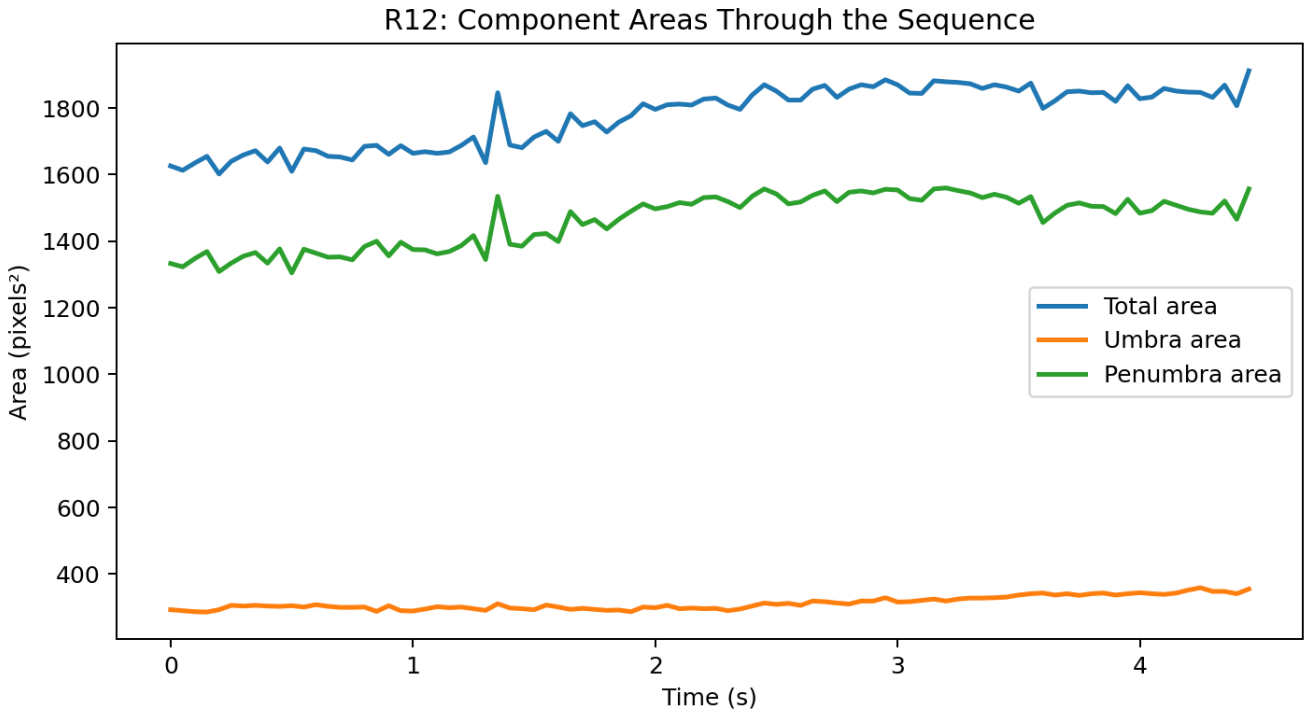
Area as a classification feature



The revised code adds Total_Area_pixels^2 and projected MSH as explicit comparison variables. MSH is a projected-area proxy and is not foreshortening-corrected.

Current measured values are from the original 90-frame sequence; the new McIntosh fields are additional outputs of the revised code.

Example of Individual-Region Evolution



R12 — largest tracked region

Observed in all 90 frames
Mean total area: 1775.4 px²
Maximum total area: 1912 px²

Component partition

Mean umbra area: 312.9 px²
Mean penumbra area: 1462.5 px²
Maximum umbra area: 359 px²

Interpretation

The short sequence shows that the pipeline can preserve a region identity and quantify component areas frame by frame. It does not establish a physical growth or decay timescale.

Machine-Learning Stage: What Is Implemented and What Remains

Implemented in the code

RandomForestClassifier
300 trees
class_weight = balanced
min_samples_leaf = 2
GroupShuffleSplit by Region_ID when possible
ML_McIntosh_Class
ML_Confidence
ML_Model_Info Excel sheet

Training requirement

A labelled CSV is required:

Region_ID
McIntosh_Class

The model is not activated without this file.

Therefore the present run cannot legitimately be reported as a validated ML classification experiment unless labelled training data are supplied.

Validation protocol

Avoid frame leakage by splitting at Region_ID level.

Report:
- accuracy
- precision
- recall
- F1-score
- confusion matrix
- class-wise performance

Compare proxy vs. expert/catalogue labels.