Crowdsourced Keypoint Matching for Region of Interest Identification in Cross-Stain Whole Slide Image Registration

Correspondence-oriented crowdsourcing with robust filtering and gradient correlation validation for H&E-IHC ROI transfer

Alessio Fiorin*[1,2,3], Eddy Maddalena*[4], Michael Soprano*[4], Laia Adalid-Llansa[2,1], Vincenzo Della Mea[4], Carlos López-Pablo[2,1]

[1] Institut de Recerca Biomèdica Catalunya Sud, Spain · [2] Hospital Universitari de Tortosa Verge de la Cinta, Catalan Health Institute, Spain · [3] Universitat Rovira i Virgili, Spain · [4] University of Udine, Italy · * Equal contribution

Funded by the European Union

Problem Statement, Dataset, and Contribution

Introduction and Background

Digital pathology often requires transferring a region of interest (ROI) from an H&E whole-slide image (WSI) to matched IHC WSIs from consecutive sections. Such transfer enables the joint interpretation of tissue morphology and biomarker signals, such as TIL assessment on H&E and immune population characterization on IHC

- H&E and IHC are not the same physical section, so the ROI must be localized again on each IHC slide
- Cross-stain appearance shifts, weak staining, tissue loss, non-target tissue, and rotations reduce reliable visual overlap
- Manual ROI relocation is slow and error-prone; automatic feature matching may fail in out-of-distribution cases

Objective and Contribution

We investigate whether non-expert crowd workers can provide reliable keypoint correspondences for cross-stain ROI transfer. Crowd-derived correspondences are compared with classical feature-based methods and deep learning-based sparse, semi-dense, and dense matchers

- A correspondence-oriented crowdsourcing protocol for difficult H&E-IHC WSI pairs
- A unified comparison against automatic feature-matching baselines
- An analysis of structural validation through gradient correlation and worker-level transform selection

Problem Statement: ROI Transfer Task

Fixed image: F = H&E WSI; moving image: M = IHC WSI

Estimate rigid transform T from matched keypoint correspondences

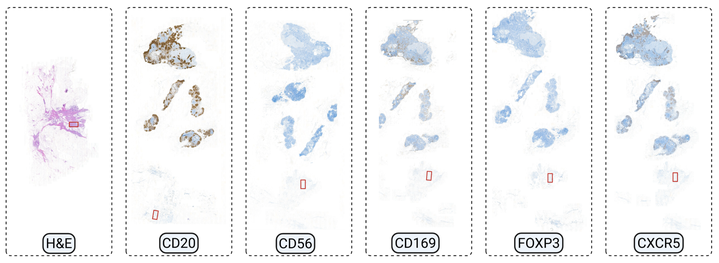
$R_F = \{p_1, p_2, p_3, p_4\} \subset F$

$r_i = T^{-1}(p_i),\ i = 1, \ldots, 4$

$R_M = \{r_1, r_2, r_3, r_4\} \subset M$

$\text{cTRE} = ||c(R_M) - c(R_M^{gt})||_2$

Dataset



Example H&E/IHC pair series with ROI boxes across biomarkers

32 H&E-IHC pairs

4 Breast cancer patients

8 IHC biomarkers

20× Scanner magnification

- The cohort includes primary tumor tissue and two axillary lymph node biopsies for each case
- The eight biomarkers are CD4, FOXP3, CXCR5, CD20, CD21, CD163, CD169, and CD56
- Several IHC slides show heterogeneous staining, weak target-to-background contrast, and tissue artifacts

Cross-Stain Challenges

Appearance shift
Color, contrast, and tissue texture change substantially between H&E and IHC

Tissue Loss
Consecutive sections may miss parts of the target tumor region

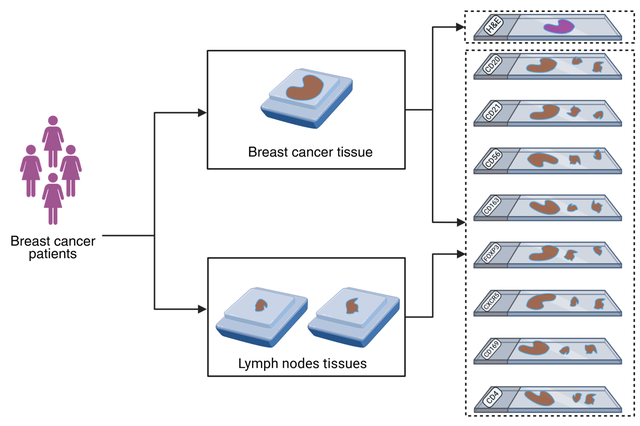
Non-target Tissue
IHC slides may include lymph node tissue that is absent from the H&E ROI context

Rotation
Orientation differences make automatic matching less reliable without explicit rotation handling

Weak Staining
Reduced target-to-background contrast makes corresponding structures harder to identify

Artifacts
Histological artifacts and weak or uneven staining can break feature matching

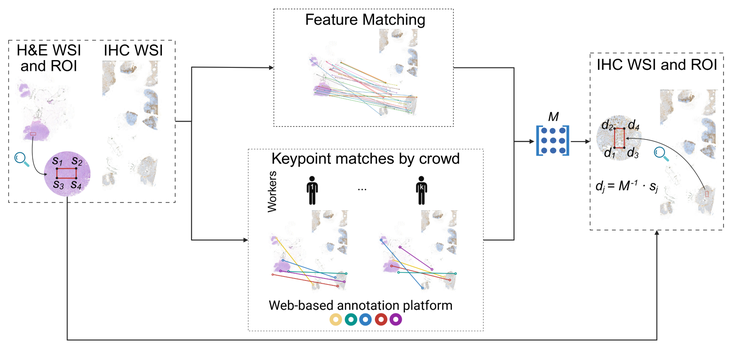
When automatic matchers return unstable or wrong correspondences, human visual matching may still exploit anatomical structures that remain recognizable across stains



Cohort structure: breast cancer patients, tissue sections, and matched WSI series

Crowdsourcing Task and Experimental Setup

Proposed Pipeline



Starting from an H&E WSI with a pathologist-annotated ROI and a corresponding IHC WSI, the pipeline obtains point correspondences, estimates a rigid transformation, validates alignment quality, and transfers the ROI to the IHC slide. Automatic methods additionally test the IHC slide under four rotations: 0°, -90°, 90°, and 180°

Automatic Feature Matching Baselines

- Classical pipelines include SIFT and BRISK with standard detection, description, and matching
- Deep and dense matchers include SP+SG, LoFTR, eLoFTR, ASpanFormer, DKM, RoMa, and MatchAnything
- RANSAC removes geometric outliers; an additional IQR filtering variant removes residual mismatches
- Rigid transformations are estimated with the Kabsch-Umeyama algorithm at downsampled WSI resolution

Crowdsourcing Task

Human keypoint correspondences were collected on Prolific with a custom web-based annotation platform. Each annotation unit contained four H&E/IHC image pairs. For each pair, workers placed five markers on H&E and five corresponding markers on IHC.

5 workers
Annotated each image pair

25 matches
Available for each pair before filtering

40 units
Completed in the final deployment

800 matches
Collected before filtering and aggregation

Crowd Aggregation Strategies

$\text{Crowd}_{\text{All}}$
Pool all 25 correspondences and apply the same robust filtering used for automatic methods

µTRE
Mean keypoint error between warped IHC and H&E crowd landmarks

LoO
Leave-one-out geometric stability of each worker transformation

GradCorr
Gradient correlation between registered enhanced H&E and IHC images

Combined Ranking
Balanced combination of structural and geometric quality criteria

Deployment Details

£0.70/unit | **£40 total** | **16.0 min med.** | **1h46**

Data collection finished in 1 hour and 46 minutes, at a total cost of £40.00 including platform fees

Crowd Selection Logic

1. Estimate one rigid transform from each worker correspondence set
2. Measure geometric consistency using µTRE and leave-one-out error
3. Measure structural agreement with GradCorr on enhanced registered images
4. Rank transformations using complementary geometric and structural evidence
5. Use the selected transform to transfer the four ROI vertices to IHC

µTRE and LoO both summarize geometric consistency, so they are not directly combined. GradCorr adds structural validation by checking whether aligned tissue structures agree

Results and Discussion

Evaluation Metrics

- ROI localization accuracy is measured using centroidTarget Registration Error (cTRE)
- cTRE is the Euclidean distance between predicted and ground-truth ROI centroids annotated by an expert pathologist
- Success rates use weak, medium, and strict empirically defined tolerance levels

Success Rates (SR)

Registration Accuracy

Feature Matching	SR_{tW}	SR_{tM}	SR_{tS}	$cTRE_{median}$	$cTRE_{IQR}$
SIFT	0.0	0.0	0.0	/	/-/
BRISK	34.4	28.1	28.1	13.182	0.911-21.267
SP+SG	43.8	43.8	37.5	5.088	0.675-15.409
LoFTR	46.9	46.9	46.9	13.455	0.502-20.520
eLoFTR	62.5	62.5	56.2	0.665	0.390-17.607
ASpanFormer	62.5	62.5	62.5	0.631	0.356-15.712
DKM	68.8	68.8	65.6	0.610	0.430-7.114
RoMa	65.6	65.6	65.6	0.719	0.400-8.397
MatchAnything$_{\text{eLoFTR}}$	43.8	40.6	31.2	14.792	0.831-19.936
MatchAnything$_{\text{RoMa}}$	68.8	68.8	68.8	**0.489**	0.344-6.939
$\text{Crowd}_{\text{All}}$	56.2	53.1	50.0	1.382	0.774-11.263
$\text{Crowd}_{\text{LoO}}$	68.8	68.8	59.4	0.979	0.612-9.578
$\text{Crowd}_{\mu\text{TRE}}$	75.0	75.0	68.8	0.958	0.642-2.747
$\text{Crowd}_{\text{GradCorr+µTRE}}$	90.6	90.6	**84.4**	0.871	**0.553-1.179**
$\text{Crowd}_{\text{GradCorr}}$	90.6	90.6	**84.4**	0.833	0.501-1.160
$\text{Crowd}_{\text{GradCorr+LoO}}$	**93.8**	**93.8**	81.2	0.815	0.501-1.197

SRtW, SRtM, and SRtS are weak, medium, and strict success rates. cTREmedian and IQR are in millimeters. Bold values indicate the best automatic method (blue) and the best crowd-based strategy (purple) within each metric.

cTRE Variability and Robustness

MatchAnything with RoMa backbone has the lowest cTRE median, whereas the strongest crowd variants show much narrower cTRE IQR intervals, suggesting better robustness across difficult cases

Key Results and Interpretation

- SIFT failed in all cases, showing the limits of handcrafted descriptors under severe cross-stain appearance changes
- Among automatic methods, MatchAnything$_{\text{RoMa}}$ achieved the lowest $cTRE_{median}$, but had wider variability
- $\text{Crowd}_{\text{GradCorr+LoO}}$ achieved the best weak and medium success rates: 93.8%
- $\text{Crowd}_{\text{GradCorr}}$ and $\text{Crowd}_{\text{GradCorr+µTRE}}$ achieved the best strict success rate: 84.4%
- GradCorr helps reject geometrically consistent but structurally wrong worker transformations

Practical Implications
The proposed workflow can be used as a human-in-the-loop support step for difficult H&E-IHC WSI pairs, especially when automatic feature matching returns too few reliable correspondences or unstable registrations

Conclusions and Future Work

- Experiments are limited to a single in-house cohort of challenging breast cancer cases,although the cohort includes cross-stain variability, artifacts, non-target tissues, andweak or uneven staining
- Future work will test public histological benchmarks such as ANHIR, HyReCo, and ACROBAT, and optimize worker guidance and selection
- Overall, crowdsourced keypoint correspondences can complement automatic matching inchallenging cross-stain WSI registration workflows

Take-home Message

Crowdsourcing provides a practical fallback for difficult H&E-IHC WSI pairs when automatic matchers fail or produce unreliable correspondences.

Higher success rates
Up to +25.0% at weak/medium and +15.6% at strict tolerance

Structural and geometric validation
GradCorr complements µTRE and LoO during crowd selection

Robust ROI transfer
Strong crowd variants keep low cTRE variability in this cohort

Acknowledgement EU MSCA BosomShield Doctoral Network HORIZON-MSCA-2021-DN-01-01 grant number 101073222. Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or European Research Executive Agency (REA). Neither the European Union nor the granting authority can be held responsible for them.