Calaix de sastre

Coeli-IA: *deep learning* al servei de les col·leccions patrimonials

Coeli-IA: *deep learning* al servei de les col·leccions patrimonials

Resum: Coeli-IA és un projecte de recerca que explora l'ús de tecnologies disruptives de catalogació automàtica i cerca multimodal de contingut en col·leccions de patrimoni cultural, mitjançant els últims avenços en aprenentatge profund (*deep learning*). A través de l'anàlisi del contingut visual de les imatges, s'ha creat un sistema per suggerir etiquetes i categories de manera automàtica a partir de fotografies de peces de museu, fet que agilitza la tasca de catalogació manual a les institucions. A més, en el marc d'aquest projecte també s'ha treballat en la identificació d'imatges duplicades i s'ha implementat un sistema de cerca multimodal d'imatges sense necessitat de catalogació prèvia, per tal d'oferir noves eines de recuperació de la informació als usuaris. El repte principal ha estat la millora de l'accés als continguts del sector GLAM (galeries, biblioteques, arxius i museus), amb l'objectiu de superar la dependència d'etiquetatge manual i la dificultat de gestionar grans volums de dades.

Paraules clau: Indexació automàtica, cerca multimodal, aprenentatge profund, *deep learning*, patrimoni cultural, visió per computador.

Marc Folia
Responsable d'I+D, Coeli Platform
marc.folia@coeli.cat

Francesc Net
Enginyer de Recerca, Universitat Autònoma de Barcelona, Centre de Visió per Computador
fnet@cvc.uab.cat

Lluís Gómez
Investigador Ramon y Cajal, Universitat Autònoma de Barcelona, Centre de Visió per Computador.
lgomez@cvc.uab.es

Pep Casals
Director, NUBILUM
pep.casals@nubilum.es

Data de recepció: 17/02/2025
Data d'acceptació: 31/08/2025

DOI
10.60940/itemn79id9900268

Coeli-IA: *deep learning* al servicio de las colecciones culturales

Resumen: Coeli-IA es un proyecto de investigación que explora el uso de tecnologías disruptivas de catalogación automática y búsqueda multimodal de contenido en colecciones de patrimonio cultural, mediante los últimos avances en aprendizaje profundo *(deep learning)*. A través del análisis del contenido visual de las imágenes, se ha creado un sistema para sugerir etiquetas y categorías de forma automática a partir de fotografías de piezas de museo, agilizando así la labor de catalogación manual en las instituciones. Además, en el marco de este proyecto también se ha trabajado en la identificación de imágenes duplicadas y se ha implementado un sistema de búsqueda multimodal de imágenes sin necesidad de previa catalogación, ofreciendo nuevas herramientas de recuperación de la información a los usuarios. El principal reto es mejorar el acceso a los contenidos del sector GLAM (galerías, bibliotecas, archivos y museos), superando la dependencia del etiquetado manual y

la dificultad de gestionar grandes volúmenes de datos.
Palabras clave: Indexación automática, búsqueda multimodal, aprendizaje profundo, deep learning, patrimonio cultural, visión por computador.

Coeli-IA: deep learning at the service of cultural collections

Abstract: Coeli-IA is a research project that explores the use of disruptive technologies for automatic cataloging and multimodal content search in cultural heritage collections, leveraging the latest advances in Deep Learning. By analyzing the visual content of the images, a system has been created to automatically suggest tags and categories from photographs of museum artifacts, thus streamlining the manual cataloging process in institutions. Furthermore, within the framework of this project, work has also been done on the identification of duplicate images and a multimodal image search system has been implemented without the need for prior cataloguing, offering new information retrieval tools to users. The main challenge is to improve access to content in the GLAM sector (galleries, libraries, archives and museums), overcoming the reliance on manual tagging and the difficulty of managing large volumes of data.
Keywords: Automatic indexing, multimodal search, deep learning, cultural heritage, computer vision.

Introducció

La gestió documental i la difusió del patrimoni cultural es basen en gran part en una metodologia de descripció i indexació a partir de paraules clau. Habitualment, aquestes paraules clau s'obtenen de tesaurus o vocabularis controlats de referència amb diferents nivells de complexitat. D'una banda, en molts casos aquestes tasques de catalogació poden ser oneroses en temps i recursos, i sovint esdevenen inviables quan es tracta de grans volums de documents. D'altra banda, l'ús de descriptors per a l'accés a la informació no està exempt de problemàtiques àmpliament conegudes en el sector, com per exemple la coherència en els criteris d'indexació, les diferències d'interpretació dels vocabularis entre usuaris experts i gran públic, o l'adaptació dels descriptors a les necessitats d'informació canviants.

En el projecte Coeli-IA ens hem volgut plantejar de quina manera les tecnologies basades en intel·ligència artificial ens poden ajudar en alguns d'aquests aspectes, tant per facilitar l'accés del públic en general als continguts culturals com per crear eines de suport als professionals del sector GLAM (de l'anglès, *galleries, libraries, archives and museums*). En particular, hem utilitzat tècniques d'aprenentatge profund (en anglès, deep learning) per entrenar algoritmes en l'execució de tasques de classificació, anotació i indexació a partir d'imatges i continguts visuals.

NUBILUM <www.nubilum.cat> és una empresa de serveis en l'àmbit del patrimoni cultural i documental constituïda l'abril del 2013. La seva missió és ajudar les organitzacions a ser més eficients, fent un ús eficaç de la gestió de la documentació i el coneixement. Integrar el coneixement, preservar-lo i fer-lo accessible és la raó de ser de les seves actuacions. Per a totes les organitzacions, el coneixement i la seva gestió són decisius per als resultats de la seva activitat. D'ençà que es van constituir ha fet diverses activitats d'R+D+I i ha desenvolupat el programari Coeli <www.coeli.cat>, una plataforma digital per difondre, documentar i gestionar les col·leccions patrimonials.

El Centre de Visió per Computador (CVC)[1] és un centre TECNIO de recerca sense ànim de lucre fundat el 1995 com a consorci entre la Generalitat de Catalunya i la Universitat Autònoma de Barcelona (UAB). La seva missió és dur a terme investigació capdavantera, d'excel·lència i d'alt impacte internacional en el camp de la visió per computador, una àrea de la intel·ligència artificial orientada a l'anàlisi i processament d'imatges. D'altra banda, també són missió del centre promoure la transferència de coneixement a la indústria i en la societat, així com preparar i formar investigadors al més alt nivell europeu. El CVC compta amb més de 130 investigadors multidisciplinaris i tècnics de diferents nacionalitats, especialistes i líders en el camp de la visió per computador. El centre té diferents línies de recerca dirigides a 4 àmbits d'aplicació de mercat, entre les quals es destaquen les indústries culturals i basades en l'experiència.

El programa d'ajuts INNOTEC de l'Agència per la Competitivitat de l'Empresa (ACCIÓ) ha permès dur a terme aquest projecte d'R+D. Per part de NUBILUM, l'objectiu ha estat ampliar i millorar l'eina Coeli amb tecnologia disruptiva pel que fa a la catalogació i cerca de continguts. Per la seva banda, l'objectiu principal del CVC ha estat avançar en el nivell de preparació de la seva tecnologia, creant un producte mínim viable que s'ha demostrat i validat en l'entorn productiu fent ús de dades reals.

1. Objectius del projecte

L'objectiu principal d'aquest projecte és adaptar les arquitectures de xarxes neuronals per a visió per computador al camp específic de

1. *CVC Centre de Visió per Computador.* https://www.cvc.uab.es/ca/ [Consulta: 20 novembre 2020].

la gestió de col·leccions de patrimoni cultural. L'ús dels últims avenços en el camp de l'aprenentatge profund (*deep learning*) tenen el potencial de millorar significativament la tecnologia de cerca actual de Coeli i incorporar funcionalitats de catalogació automàtica. En particular els objectius específics del projecte han estat:

— Crear un corpus d'imatges i anotacions per construir una base de coneixement en indexació automàtica.
— Avançar el nivell de maduresa de la tecnologia (TRL, de l'anglès *technology readiness level*) desenvolupada al CVC. Demostrar i validar la tecnologia en entorn productiu.
— Optimitzar els processos de treball i millorar la productivitat dels usuaris de la plataforma Coeli.
— Simplificar i facilitar la cerca a les col·leccions per als visitants digitals.

2. Eines i metodologies

El desenvolupament dels reptes tecnològics plantejats en aquest projecte ha requerit l'ús de diferents tècniques i metodologies que permeten resoldre diferents problemàtiques a partir d'un marc tecnològic comú, la visió per computador i l'aprenentatge profund o *deep learning*.

2.1 Visió per computador i deep learning

En els darrers anys, amb l'explosió de les tècniques d'aprenentatge profund tot el camp d'investigació en visió per computador s'ha revolucionat i ha assolit rendiments excepcionals en tasques que fins recentment eren considerades massa complexes d'automatitzar.[2,3] Concretament, l'ús d'aquesta tecnologia ha produït algorismes capaços de superar el rendiment humà en la tasca bàsica de classificació d'imatges[4] —una fita que era inimaginable fa una dècada— i ha obert la porta a resoldre tasques encara més complexes, que requereixen el reconeixement d'una major varietat de contingut visual, com ara l'etiquetatge i subtitulació automàtica d'imatges en llenguatge natural,[5,6] la comprensió d'escenes,[7,8] la cerca semàntica,[9] etc.

2.2 Catalogació automàtica d'imatges

La tasca d'etiquetatge o catalogació d'imatges consisteix a assignar una llista d'etiquetes a les imatges, fent referència a paraules que descriuen el contingut o el context de la imatge, però utilitzant un vocabulari de termes molt gran (desenes o fins i tot centenars de milers) i on una mateixa imatge pertany normalment a més d'una classe.

2. LeCun, Yann; Bengio, Yoshua; Hinton, Geoffrey. «Deep learning». *Nature*, 2015, vol. 521, núm. 7553, p. 436-444. DOI: 10.1038/nature14539.
3. Goodfellow, Ian; Bengio, Yoshua; Courville, Aaron. *Deep Learning: Adaptive Computation and Machine Learning series*. Cambridge: The MIT Press, 2016. ISBN: 9780262035613.
4. He, Kaiming; Zhang, Xiangyu; Ren, Shaoqing; Sun, Jian. «Delving deep into rectifiers: Surpassing human-level performance on ImageNet classification». *Proc. of the IEEE International Conference on Computer Vision (ICCV)*, 2015, p. 1026-1034.
5. Fu, Jianlong; Rui, Yong. «Advances in deep learning approaches for image tagging». *Transactions on Signal and Information Processing*, 2017, vol. 6.
6. Vinyals, Oriol; Toshev, Alexander; Bengio, Samy; Erhan, Dumitru. «Show and tell: A neural image caption generator». *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, p. 3156-3164.
7. Zhou, Bolei; Lapedriza, Agata; Khosla, Aditya; Oliva, Aude; Torralba, Antonio. «Learning deep features for scene recognition using Places database». *Advances in Neural Information Processing Systems (NIPS)*, 2014, p. 487-495.
8. Zhou, Bolei; Lapedriza, Agata; Khosla, Aditya; Oliva, Aude; Torralba, Antonio. «Places: A 10 million image database for scene recognition». *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017.
9. Gordo, Albert; Almazán, Jon; Revaud, Jérôme; Larlus, Diane. «Deep image retrieval: Learning global representations for image search». *Proceedings of the European Conference on Computer Vision (ECCV)*, 2016, p. 241-257.

Les xarxes neuronals artificials són un tipus de tècnica d'aprenentatge automàtic que permet als ordinadors reconèixer patrons i prendre decisions a partir de grans quantitats de dades. S'anomenen «neuronals» perquè la seva estructura s'inspira, de manera molt simplificada, en la manera com les neurones transmeten informació dins del sistema nerviós humà. El mètode més comú per abordar l'etiquetatge automàtic d'imatges en aquests termes consisteix a utilitzar una xarxa neuronal convolucional profunda de l'estat de l'art[10] en combinació amb una capa de regressió logística.[11] Més recentment s'han proposat nous mètodes que utilitzen millor les dependències estadístiques de coocurrència de les etiquetes.[12,13] Un algorisme desenvolupat recentment al CVC[14] permet predir etiquetes més ben contextualitzades a partir d'alguna informació complementària a la imatge (per exemple la seva geolocalització).

Les xarxes neuronals artificials són un tipus de tècnica d'aprenentatge automàtic que permet als ordinadors reconèixer patrons i prendre decisions a partir de grans quantitats de dades.

2.3. Cerca multimodal en col·leccions d'imatges

Un enfocament comú per abordar la cerca multimodal en col·leccions d'imatges és aprendre un espai conjunt de projecció (en anglès, *embedding space*) d'imatges i paraules.[15,16,17] En aquest espai, les imatges es projecten a prop de les paraules amb què comparteixen semàntica. En conseqüència, també les imatges similars es projecten en punts similars de l'espai de projecció. D'aquesta manera la cerca d'imatges a partir d'una imatge o text de referència es converteix en un simple problema de trobar-ne els veïns més propers (KNN per les seves sigles en anglès) en aquest espai de projecció. Aquest espai de projecció es pot també modelar a partir de l'agregació de descriptors d'imatge complementaris[18] en cas que l'aplicació requereixi trobar una peça concreta prèviament inventariada a partir d'una imatge proporcionada per l'usuari.

Una línia de recerca desenvolupada al CVC[19,20,21] explora la idea de projectar imatges i text en un espai comú basant-se en una col·lecció de documents multimodals (text i

10. He, Kaiming; Zhang, Xiangyu; Ren, Shaoqing; Sun, Jian. "Delving deep into rectifiers: Surpassing human-level performance on ImageNet classification". *Proc. of the IEEE International Conference on Computer Vision (ICCV)*, 2015, p. 1026-1034.
11. Fu, Jianlong; Rui, Yong. «Advances in deep learning approaches for image tagging». *Transactions on Signal and Information Processing*, 2017, vol. 6.
12. Chollet, François. «Information-theoretical label embeddings for large-scale image classification». *arXiv preprint*, 2016. https://arxiv.org/abs/1607.05690 [Consulta: 20 novembre 2020].
13. Mineiro, Paul; Karampatziakis, Nikos. «Fast label embeddings via randomized linear algebra». A: *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, Cham, 2015, p. 37-51.
14. Gómez, Raúl; Gibert, Jaume; Karatzas, Dimosthenis. «Location sensitive image retrieval and tagging». A: *European Conference on Computer Vision*. Springer, Cham, 2020, p. 401-417.
15. Akata, Zeynep; Perronnin, Florent; Harchaoui, Zaid; Schmid, Cordelia. «Label-embedding for image classification». *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2015, vol. 38, núm. 7, p. 1425-1438.
16. Frome, Andrea; Corrado, Greg S.; Shlens, Jon; Bengio, Samy; Dean, Jeff; Mikolov, Tomas. "Devise: A deep visual-semantic embedding model". *Advances in Neural Information Processing Systems*, 2013, p. 2121-2129.
17. Wang, Jiang; Yang, Yi; Mao, Junhua; Huang, Zhiheng; Huang, Chang; Xu, Wei. «Cnn-rnn: A unified framework for multi-label image classification». A: *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, p. 2285-2294.
18. Husain, Syed Sameed; Bober, Miroslaw. «REMAP: Multi-layer entropy-guided pooling of dense CNN features for image retrieval». *IEEE Transactions on Image Processing*, 2019, vol. 28, núm. 10, p. 5201-5213.
19. Gómez, Lluís; Patel, Yash; Rusinol, Marçal; Karatzas, Dimosthenis. «Self-supervised learning of visual features through embedding images into text topic spaces». A: *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, p. 4230-4239.
20. Gómez, Raúl; Gibert, Jaume; Karatzas, Dimosthenis. Self-supervised learning from web data for multimodal retrieval. A: Multimodal Scene Understanding. Academic Press, 2019, p. 279-306.
21. Gómez, Raúl; Gibert, Jaume; Karatzas, Dimosthenis. «Location sensitive image retrieval and tagging». A: *European Conference on Computer Vision*. Springer, Cham, 2020, p. 401-417.

imatges). Per als objectius d'aquest projecte, es pot aprofitar un marc similar per establir corelacions semàntiques contextuals entre les imatges proporcionades pels usuaris i el material existent en col·leccions de les institucions que usen Coeli.

3. Reptes tecnològics del projecte

Un dels principals reptes tecnològics d'aquest projecte té a veure amb l'ús de vocabularis amb un gran nombre d'etiquetes. Un dels vocabularis controlats més difusos en la gestió de col·leccions a Catalunya és l'Art & Architecture Thesaurus,[22] que disposa d'una traducció catalana realitzada al Departament de Cultura de la Generalitat de Catalunya[23] i que conté més de 165.000 termes. Un repte important en aquest escenari és el de reconèixer amb èxit les classes o conceptes menys freqüents, el que estadísticament es coneix com la llarga cua d'una distribució de dades, en la base de dades d'entrenament, mitjançant la combinació de tècniques de *zero-shot learning*[24,25] amb les arquitectures d'etiquetatge existents.

D'altra banda, la supervisió d'algorismes d'aprenentatge automàtic a partir d'etiquetes comporta un conjunt de reptes propis. El més conegut és el problema de les anotacions incompletes,[26] que es produeix pel fet que dos anotadors humans poden considerar rellevants diferents subconjunts d'etiquetes per a una mateixa imatge, malgrat que ambdues anotacions poden ser igualment vàlides. Això dificulta l'aprenentatge perquè el model rep diferents informacions per a imatges similars.

Per això hem desenvolupat un sistema basat en intel·ligència artificial capaç de suggerir etiquetes per a objectes culturals a partir de la seva representació visual, utilitzant xarxes neuronals profundes entrenades amb conjunts de dades reals i estructures d'etiquetatge jeràrquiques.

Un altre repte important és el de gestionar la incertesa dels models d'etiquetatge i cerca. Mentre que en proves de laboratori els errors dels models són simplement un valor numèric que s'ha d'intentar minimitzar, en un entorn comercial els errors poden desencoratjar l'usuari i dificultar el desplegament de l'eina en alguns entorns operacionals reals. Per evitar aquestes situacions s'hauran d'integrar tècniques per quantificar la incertesa del model[27] i utilitzar-les com a part dels models de predicció per millorar l'experiència de l'usuari.

3.1 Anotació i categorització automàtica

L'anotació automàtica d'imatges en col·leccions patrimonials suposa un repte considerable per la seva diversitat semàntica i la

22. *Getty Research Institute. Art & Architecture Thesaurus® Online*. https://www.getty.edu/research/tools/vocabularies/aat/ [Consulta: 20 novembre 2020].
23. *Tesaurus d'Art i Arquitectura*. https://aatesaurus.cultura.gencat.cat/index.php [Consulta: 20 novembre 2020].
24. Norouzi, Mohammad; Mikolov, Tomas; Bengio, Samy; Singer, Yoram; Shlens, Jonathon; Frome, Andrea; Corrado, Greg S.; Dean, Jeffrey. «Zero-shot learning by convex combination of semantic embeddings». A: *2nd International Conference on Learning Representations (ICLR)*, 2014.
25. Zhang, Yang; Gong, Boqing; Shah, Mubarak. «Fast zero-shot image tagging». A: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.
26. Wang, Qifan; Shen, Bin; Wang, Shumiao; Li, Liang; Si, Luo. «Binary codes embedding for fast image tagging with incomplete labels». A: *European Conference on Computer Vision*. Springer, Cham, 2014, p. 425-439.
27. Lakshminarayanan, Balaji; Pritzel, Alexander; Blundell, Charles. «Simple and scalable predictive uncertainty estimation using deep ensembles». *Advances in neural information processing systems*, 2017, vol. 30.

complexitat dels esquemes de classificació utilitzats en el sector GLAM. Per això hem desenvolupat un sistema basat en intel·li gència artificial capaç de suggerir etiquetes per a objectes culturals a partir de la seva representació visual, utilitzant xarxes neuronals profundes entrenades amb conjunts de dades reals i estructures d'etiquetatge jeràrquiques pròpies del sector.

Per al desenvolupament i validació del sistema, s'han utilitzat dos conjunts de dades principals, l'EUFCC-340K,[28] un corpus públic creat a partir de dades del portal Europeana <www.europeana.eu>, format per més de 340.000 imatges d'objectes patrimonials, i el corpus Coeli, que conté materials procedents d'algunes col·leccions gestionades a través d'aquesta plataforma. Cal destacar que en ambdós casos es tracta de fons d'imatges característics de l'àmbit patrimonial, en particular d'institucions museístiques, i que, per tant, han permès treballar amb major precisió les necessitats específiques del sector.

Els dos conjunts segueixen un esquema d'anotació basat en l'esmentat Art & Architecture Thesaurus (AAT). L'AAT organitza els termes en facetes que representen diferents aspectes dels objectes catalogats. En el nostre sistema, ens hem centrat en les següents facetes principals:

— Tipus d'objecte: classifica els elements segons la seva funció o tipologia.
— Materials: defineix els components físics d'un objecte.
— Disciplines: relaciona l'objecte amb camps del coneixement.
— Temes: categoritza els objectes segons el seu contingut iconogràfic o semàntic.

L'estructura jeràrquica de l'AAT permet assignar etiquetes en diferents nivells de granularitat que faciliten la inferència de categories generals en casos on no hi ha prou informació per a una classificació específica. Per exemple, si el sistema detecta una escultura metàl·lica però no pot determinar-ne el tipus

Figura 1: Exemple del prototip inicial de classificació automàtica.

28. Net, Francesc; Folia, Marc; Casals, Pep; Bagdanov, Andrew D.; Gómez, Lluís. «EUFCC-340K: A faceted hierarchical dataset for metadata annotation in GLAM collections». *Multimedia Tools and Applications* (2025).

La gestió d'arxius patrimonials digitals requereix eines eficients per recuperar contingut visual sense dependre exclusivament de metadades textuals.

exacte, pot assignar l'etiqueta «escultura» i «metall» com a categories superiors dins la jerarquia.

Per abordar aquesta classificació complexa, s'han entrenat diversos models basats en les xarxes neuronals ConvNeXT[29] i CLIP,[30] que s'han adaptat per a tasques de classificació multietiqueta. Aquests models han estat dissenyats per gestionar la simultaneïtat de múltiples categories en una mateixa imatge i aprofitar l'organització jeràrquica de les etiquetes per millorar la precisió en categories menys freqüents. S'han provat diferents estratègies per millorar-ne el rendiment i se n'ha fet una anàlisi comparativa.

Els experiments realitzats tant en el conjunt de dades EUFCC-340K com en el corpus de Coeli han demostrat que aquest enfocament permet en molts casos obtenir una precisió elevada en l'anotació automàtica, fet que permet reduir el temps dedicat a la catalogació manual. A més, el sistema inclou un mecanisme per quantificar el grau de confiança en cada predicció i permet als professionals validar i corregir les etiquetes de manera eficient.

3.2 Cerca intel·ligent d'imatges

La gestió d'arxius patrimonials digitals requereix eines eficients per recuperar contingut visual sense dependre exclusivament de metadades textuals. En el marc del projecte Coeli-IA hem desenvolupat i integrat dues tecnologies avançades de cerca en col·leccions d'imatges: detecció de duplicats i imatges similars,[31] i cerca multimodal basada en text i imatge.[32] Aquests dos enfocaments complementaris faciliten tant la gestió interna de les col·leccions com l'experiència de cerca dels usuaris.

3.2.1 Detecció de duplicats i imatges similars

Els fons fotogràfics sovint contenen imatges duplicades o quasi idèntiques, ja sigui per repeticions en la digitalització de materials físics, seqüències de captures d'una mateixa escena o variacions en la qualitat i edició de les imatges. Identificar aquestes duplicacions de manera automàtica és clau per evitar dedicar esforços innecessaris en la catalogació i garantir una gestió més eficient dels fons documentals.

En aquest projecte hem implementat un sistema de detecció d'imatges duplicades i quasi-duplicades basat en xarxes neuronals profundes, utilitzant un enfocament d'aprenentatge transductiu.[33] Aquesta metodologia ens ha permès afinar els models directament sobre les col·leccions de Coeli, aprofitant tècniques d'autoaprenentatge[34] per detectar patrons de

29. Liu, Zhuang; Mao, Hanzi; Wu, Chao-Yuan; Feichtenhofer, Christoph; Darrell, Trevor; Xie, Saining. «A ConvNet for the 2020s». *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, p. 11976-11986.
30. Radford, Alec; Kim, Jong Wook; Hallacy, Chris; Ramesh, Aditya; Goh, Gabriel; Agarwal, Sandhini; Sastry, Girish; Askell, Amanda; Mishkin, Pamela; Clark, Jack; Krueger, Gretchen; Sutskever, Ilya. «Learning transferable visual models from natural language supervision». A: *Proceedings of the 38th International Conference on Machine Learning*, 2021, p. 8748-8763.
31. Net, Francesc; Folia, Marc; Casals, Pep; Gómez, Lluís. «Transductive Learning for Near-Duplicate Image Detection in Scanned Photo Collections». A: *Proc. Document Analysis and Recognition - ICDAR 2023*.
32. Zapatero, Enric; Gumà, Xavier; Net, Francesc; Gómez, Lluís; Folia, Marc. «Indexació automàtica de fotografies en un departament de comunicació: el cas de l'Institut Municipal de Mercats de Barcelona». A: *18th Image & Research Conference*, 2024.
33. Net, Francesc; Folia, Marc; Casals, Pep; Gómez, Lluís. «Transductive Learning for Near-Duplicate Image Detection in Scanned Photo Collections». A: *Proc. Document Analysis and Recognition - ICDAR 2023*.
34. Chen, Ting; Kornblith, Simon; Norouzi, Mohammad; Hinton, Geoffrey. «A simple framework for contrastive learning of visual representations». A: *Proceedings of the 37th International Conference on Machine Learning*, 2020, p. 1597-1607.

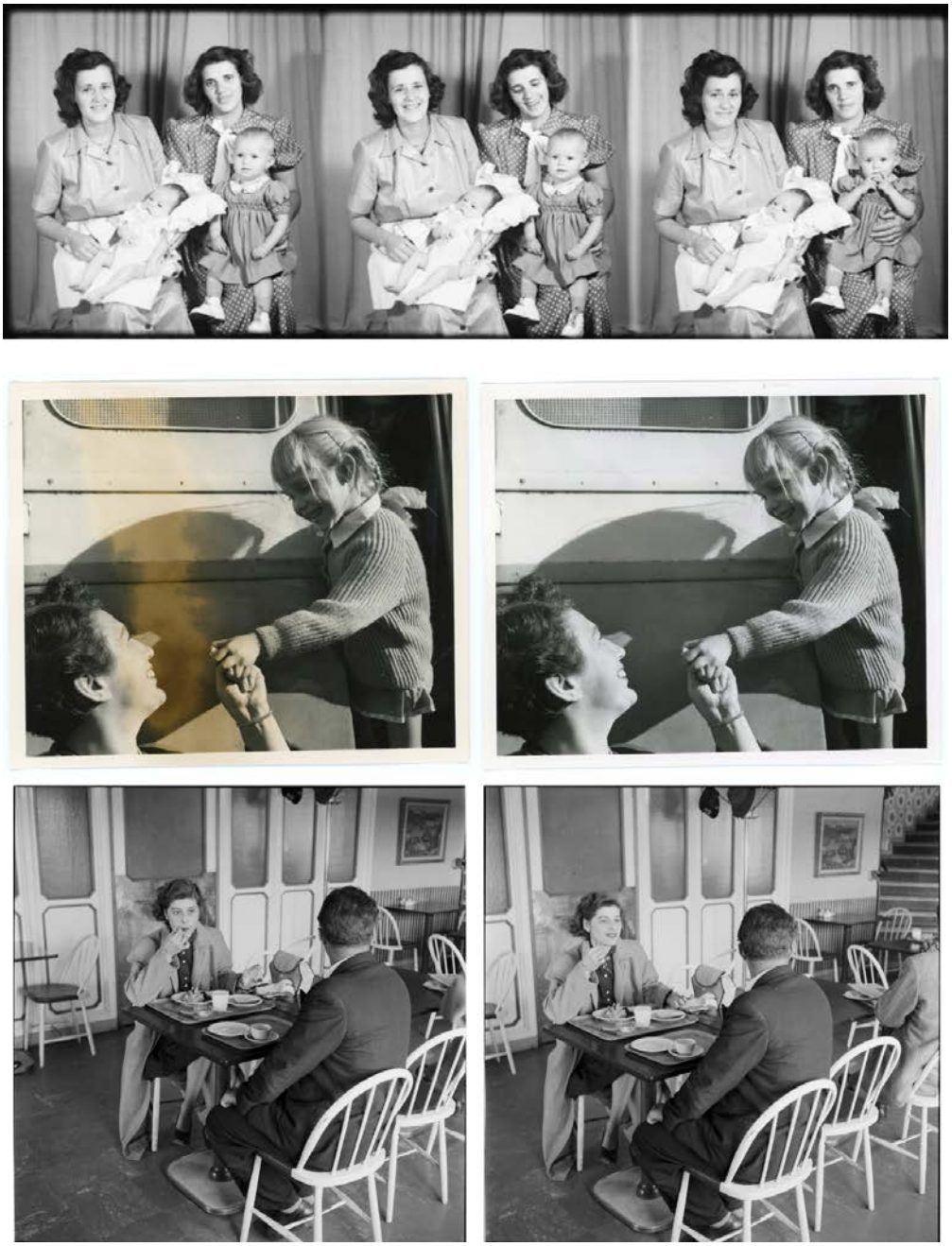

Figura 2: Exemples de duplicats o quasiduplicats en una col·lecció fotogràfica per diferents motius: (a) fotografies de la mateixa sessió d'estudi, (b) imatges escanejades de dos originals en paper amb lleugeres variacions a causa del desgast del paper al llarg del temps, (c) plans diferents en una seqüència de la mateixa escena, etc.[35]

35. Les imatges d'aquesta figura provenen del DigitaltMuseum, amb llicències Creative Common (CC BY-NC-ND i CC Public Domain). Els autors respectius són: (a) Carl Johansson, (b) Anna Riwkin-Brick i (c) Sune Sundahl

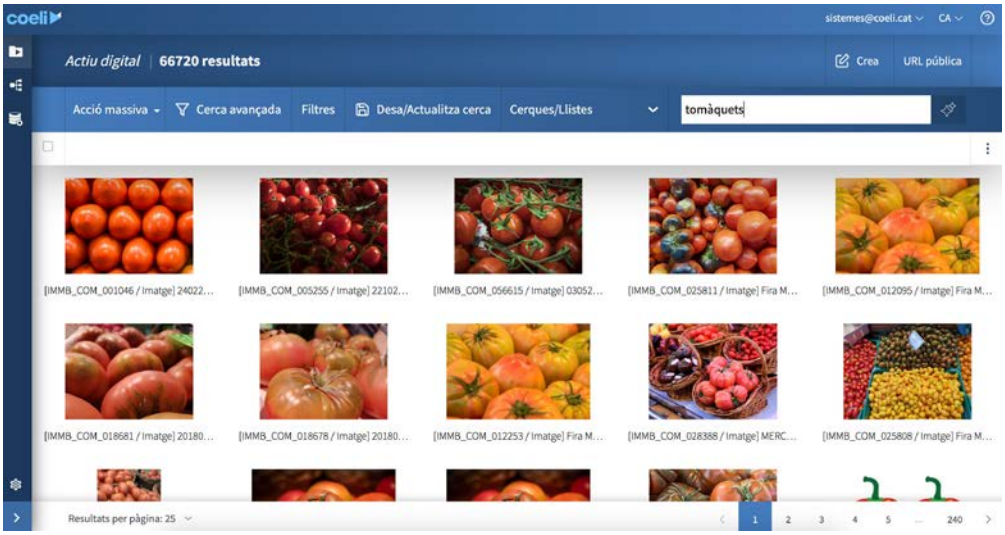

Figura 3: Funcionalitat de cerca intel·ligent de fotografies sense necessitat d'etiquetatge.

similitud sense necessitat d'anotacions manuals. Els models utilitzats es basen en les arquitectures de xarxes neuronals ResNet[36] i Vision Transformers (ViTs),[37] que generen representacions numèriques (*embeddings*) de les imatges i les comparen mitjançant mètriques de semblança.

Els experiments fets sobre el conjunt de dades públic UKBench[38] i el corpus de Coeli han mostrat que aquest sistema pot detectar duplicats amb una precisió elevada, fet que millora els mètodes tradicionals, fins i tot en casos de variacions lleus (angles diferents, il·luminació, retocs digitals, etc.). Aquesta eina s'ha integrat dins de la plataforma Coeli i permet als professionals del sector identificar i gestionar continguts redundants amb eficàcia.

3.2.2 Cerca multimodal: text i imatge

Una de les limitacions dels sistemes tradicionals de recuperació d'informació en fons fotogràfics és que depenen exclusivament de metadades textuals. Aquesta limitació es fa especialment evident en entorns on es generen grans volums d'imatges de manera contínua, com ara els departaments de comunicació d'empreses i institucions. Aquests departaments acumulen col·leccions visuals massives —reportatges fotogràfics, material per a xarxes socials, documentació d'esdeveniments, etc.— que sovint no es cataloguen de manera individualitzada, fet que en dificulta la reutilització i consulta posterior. En aquests contextos, una cerca eficaç basada en contingut visual esdevé fonamental per optimitzar la gestió d'aquestes col·leccions.

36. He, Kaiming; Zhang, Xiangyu; Ren, Shaoqing; Sun, Jian. «Deep residual learning for image recognition». A: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, p. 770-778.
37. Dosovitskiy, Alexey; Beyer, Lucas; Kolesnikov, Alexander; Weissenborn, Dirk; Zhai, Xiaohua; Unterthiner, Thomas; Dehghani, Mostafa; Minderer, Matthias; Heigold, Georg; Gelly, Sylvain; Uszkoreit, Jakob; Houlsby, Neil. «An image is worth 16x16 words: Transformers for image recognition at scale». A: *International Conference on Learning Representations (ICLR)*, 2021.
38. Nister, David; Stewenius, Henrik. «Scalable recognition with a vocabulary tree». A: *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2006. Vol. 2.

Una de les limitacions dels sistemes tradicionals de recuperació d'informació en fons fotogràfics és que depenen exclusivament de metadades textuals.

Per superar aquest obstacle, hem implementat un sistema de cerca multimodal, que permet als usuaris trobar imatges a partir d'una descripció en llenguatge natural, sense necessitat que les imatges hagin estat prèviament etiquetades.

Aquest sistema es basa en el model CLIP (de l'anglès *contrastive language-image pretraining*), que aprèn a relacionar imatges i textos en un espai semàntic comú. Quan un usuari introdueix una cerca textual, el sistema converteix aquesta consulta en una representació numèrica i la compara amb les representacions de les imatges indexades, per retornar les més similars. Això permet cerques més flexibles i intuïtives, com ara: «cadira de fusta antiga», «pintures religioses del segle XVIII», «fotografia de noia tocant la guitarra dalt d'un escenari», «flors de color groc», etc. i pot recuperar imatges que responen a la necessitat d'informació expressada en la cerca sense necessitat d'una indexació textual prèvia dels continguts.

Aquest sistema de cerca multimodal d'imatges ha estat implementat en un entorn de producció sobre un fons de més de 50.000 imatges d'un departament de comunicació i ha estat avaluat juntament amb els usuaris finals.[39] L'experiència ha demostrat que aquest sistema millora l'accés a les col·leccions digitals, especialment en aquells casos en què els usuaris no coneixen exactament les etiquetes utilitzades en la catalogació tradicional. En institucions amb volums molt grans de material fotogràfic, la cerca multimodal facilita la identificació ràpida d'imatges rellevants sense requerir processos d'indexació manuals intensius.

4. Conclusions

Les fotografies i els materials multimèdia són elements cada vegada més presents en la gestió del patrimoni cultural. La facilitat amb què produïm aquests materials i la seva qualitat fa que adquireixin un paper cada vegada més preponderant en la gestió i difusió del patrimoni. Alhora, però, la seva abundància també suposa un repte per a la gestió d'aquests mateixos materials.

L'auge, en els darrers anys, de l'anomenada intel·ligència artificial ha portat a la llum un bon nombre d'eines i tècniques basades en arquitectures de xarxes neuronals que ens permeten realitzar tasques de tractament dels fons audiovisuals que fins ara crèiem inassequibles. El repte d'aquest projecte ha estat precisament acostar aquestes tecnologies a les pràctiques i necessitats del sector GLAM, i ho hem fet emprant, per a l'entrenament dels models, fons d'imatges específics de patrimoni cultural, i adaptant els processos de reconeixement i classificació als criteris documentals i vocabularis del sector.

Algunes d'aquestes tècniques ens permeten optimitzar o repensar determinats fluxos de treball en la gestió de col·leccions. Altres ens permeten automatitzar processos de tractament d'imatges que en molts casos hauríem descartat per massa onerosos. I encara d'altres ens plantegen escenaris realment disruptius amb les pràctiques documentals consuetudinàries. Sigui com sigui, la tecnologia ens planteja constantment el repte de repensar les nostres eines i les nostres metodologies de treball per tal d'empènyer els límits d'allò possible cada dia una mica més enllà.

39. Zapatero, Enric; Gumà, Xavier; Net, Francesc; Gómez, Lluís; Folia, Marc. «Indexació automàtica de fotografies en un departament de comunicació: el cas de l'Institut Municipal de Mercats de Barcelona». A: *18th Image & Research Conference*, 2024.

Bibliografia

Akata, Zeynep; Perronnin, Florent; Harchaoui, Zaid; Schmid, Cordelia. «Label-embedding for image classification». *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2015, vol. 38, núm. 7, p. 1425-1438.

Chen, Ting; Kornblith, Simon; Norouzi, Mohammad; Hinton, Geoffrey. «A simple framework for contrastive learning of visual representations». A: *Proceedings of the 37th International Conference on Machine Learning*, 2020, p. 1597-1607.

Chollet, François. «Information-theoretical label embeddings for large-scale image classification». *arXiv preprint*, 2016. https://arxiv.org/abs/1607.05690 [Consulta: 20 novembre 2020].

CVC Centre de Visió per Computador. https://www.cvc.uab.es/ca/ [Consulta: 20 novembre 2020].

Dosovitskiy, Alexey; Beyer, Lucas; Kolesnikov, Alexander; Weissenborn, Dirk; Zhai, Xiaohua; Unterthiner, Thomas; Dehghani, Mostafa; Minderer, Matthias; Heigold, Georg; Gelly, Sylvain; Uszkoreit, Jakob; Houlsby, Neil. «An image is worth 16x16 words: Transformers for image recognition at scale». A: *International Conference on Learning Representations (ICLR)*, 2021.

Frome, Andrea; Corrado, Greg S.; Shlens, Jon; Bengio, Samy; Dean, Jeff; Mikolov, Tomas. «Devise: A deep visual-semantic embedding model». *Advances in Neural Information Processing Systems*, 2013, p. 2121-2129.

Fu, Jianlong; Rui, Yong. «Advances in deep learning approaches for image tagging». *Transactions on Signal and Information Processing*, 2017, vol. 6.

Getty Research Institute. Art & Architecture Thesaurus® Online. https://www.getty.edu/research/tools/vocabularies/aat/ [Consulta: 20 novembre 2020].

Gómez, Lluís; Patel, Yash; Rusinol, Marçal; Karatzas, Dimosthenis. «Self-supervised learning of visual features through embedding images into text topic spaces». A: *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, p. 4230-4239.

Gómez, Raúl; Gibert, Jaume; Karatzas, Dimosthenis. «Location sensitive image retrieval and tagging». A: *European Conference on Computer Vision*. Springer, Cham, 2020, p. 401-417.

Gómez, Raúl; Gibert, Jaume; Karatzas, Dimosthenis. «Self-supervised learning from web data for multimodal retrieval». A: Multimodal Scene Understanding. Academic Press, 2019, p. 279-306.

Goodfellow, Ian; Bengio, Yoshua; Courville, Aaron. *Deep Learning: Adaptive Computation and Machine Learning series*. Cambridge: The MIT Press, 2016. ISBN: 9780262035613.

Gordo, Albert; Almazán, Jon; Revaud, Jérôme; Larlus, Diane. «Deep image retrieval: Learning global representations for image search». *Proceedings of the European Conference on Computer Vision (ECCV)*, 2016, p. 241-257.

He, Kaiming; Zhang, Xiangyu; Ren, Shaoqing; Sun, Jian. «Delving deep into rectifiers: Surpassing human-level performance on ImageNet classification». *Proc. of the IEEE International Conference on Computer Vision (ICCV)*, 2015, p. 1026-1034.

He, Kaiming; Zhang, Xiangyu; Ren, Shaoqing; Sun, Jian. «Deep residual learning for image recognition». A: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, p. 770-778.

Husain, Syed Sameed; Bober, Miroslaw. «REMAP: Multi-layer entropy-guided pooling of dense CNN features for image retrieval». *IEEE Transactions on Image Processing*, 2019, vol. 28, núm. 10, p. 5201-5213.

Lakshminarayanan, Balaji; Pritzel, Alexander; Blundell, Charles. «Simple and scalable predictive uncertainty estimation using deep ensembles». *Advances in neural information processing systems*, 2017, vol. 30.

LeCun, Yann; Bengio, Yoshua; Hinton, Geoffrey. «Deep learning». *Nature*, 2015, vol. 521, núm. 7553, p. 436-444. DOI: 10.1038/nature14539.

Liu, Zhuang; Mao, Hanzi; Wu, Chao-Yuan; Feichtenhofer, Christoph; Darrell, Trevor; Xie, Saining. «A ConvNet for the 2020s». *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, p. 11976-11986.

Mineiro, Paul; Karampatziakis, Nikos. «Fast label embeddings via randomized linear algebra». A: *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, Cham, 2015, p. 37-51.

Net, Francesc; Folia, Marc; Casals, Pep; Gómez, Lluís. «Transductive Learning for Near-Duplicate Image Detection in Scanned Photo Collections». A: *Proc. Document Analysis and Recognition - ICDAR 2023.*

Net, Francesc; Folia, Marc; Casals, Pep; Bagdanov, Andrew D.; Gómez, Lluís. «EUFCC-340K: A faceted hierarchical dataset for metadata annotation in GLAM collections». *Multimedia Tools and Applications* (2025).

Nister, David; Stewenius, Henrik. «Scalable recognition with a vocabulary tree». A: *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2006. Vol. 2.

Norouzi, Mohammad; Mikolov, Tomas; Bengio, Samy; Singer, Yoram; Shlens, Jonathon; Frome, Andrea; Corrado, Greg S.; Dean, Jeffrey. «Zero-shot learning by convex combination of semantic embeddings». A: *2nd International Conference on Learning Representations (ICLR)*, 2014.

Radford, Alec; Kim, Jong Wook; Hallacy, Chris; Ramesh, Aditya; Goh, Gabriel; Agarwal, Sandhini; Sastry, Girish; Askell, Amanda; Mishkin, Pamela; Clark, Jack; Krueger, Gretchen; Sutskever, Ilya. «Learning transferable visual models from natural language supervision». A: *Proceedings of the 38th International Conference on Machine Learning*, 2021, p. 8748-8763.

Tesaurus d'Art i Arquitectura. https://aatesaurus.cultura.gencat.cat/index.php [Consulta: 20 novembre 2020].

Vinyals, Oriol; Toshev, Alexander; Bengio, Samy; Erhan, Dumitru. «Show and tell: A neural image caption generator». *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, p. 3156-3164.

Wang, Jiang; Yang, Yi; Mao, Junhua; Huang, Zhiheng; Huang, Chang; Xu, Wei. «Cnn-rnn: A unified framework for multi-label image classification». A: *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, p. 2285-2294.

Wang, Qifan; Shen, Bin; Wang, Shumiao; Li, Liang; Si, Luo. «Binary codes embedding for fast image tagging with incomplete labels». A: *European Conference on Computer Vision*. Springer, Cham, 2014, p. 425-439.

Zapatero, Enric; Gumà, Xavier; Net, Francesc; Gómez, Lluís; Folia, Marc. «Indexació automàtica de fotografies en un departament de comunicació: el cas de l'Institut Municipal de Mercats de Barcelona». A: *18th Image & Research Conference*, 2024.

Zhang, Yang; Gong, Boqing; Shah, Mubarak. «Fast zero-shot image tagging». A: *IEEE Conference on Computer Vision and Pattern Recognition* (CVPR), 2016.

Zhou, Bolei; Lapedriza, Agata; Khosla, Aditya; Oliva, Aude; Torralba, Antonio. «Learning deep features for scene recognition using Places database». *Advances in Neural Information Processing Systems (NIPS)*, 2014, p. 487-495.

Zhou, Bolei; Lapedriza, Agata; Khosla, Aditya; Oliva, Aude; Torralba, Antonio. «Places: A 10 million image database for scene recognition». *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017. ■