BULETINUL INSTITUTULUI POLITEHNIC DIN IAŞI
Publicat de
Universitatea Tehnică „Gheorghe Asachi” din Iaşi
Volumul 68 (72), Numărul 2, 2022
Secţia
ELECTROTEHNICĂ. ENERGETICĂ. ELECTRONICĂ
DOI:10.2478/bipie-2022-0009

AUTOMATIC REAL-ESTATE IMAGE ANALYSIS FOR RETRIEVAL AND CLASSIFICATION

BY

CORNELIU FLOREA[1,2,*], ANDREI RACOVIȚEANU[1,2], LAURA FLOREA[1,2] and BOGDAN FLOREA[1,2]

[1]Image Processing and Analysis Laboratory (LAPI), University Politehnica of București (UPB)
[2]SC Imopedia SRL

Received: February 9, 2022
Accepted for publication: December 10, 2022

Abstract. Nowadays, the real estate sector is also benefiting from the advance of the technology. Digital images accompany announcements advertising for renting or selling a property. Computer vision and image analysis contribute to the understanding of these images towards better consumer experience. In this paper we show how retrieval and description can be used efficiently to isolate cvasi-duplicate images and identify similar examples to the ones that have drawn viewer’s interest. Three problems related to real-estate imaging are identified and solutions are proposed and evaluated on large databases.

Keywords: image description, image retrieval, embeddings, classification.

*Corresponding author; *e-mail*: corneliu.florea@upb.ro
© 2022 Corneliu Florea et al.
This is an open access article licensed under the Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License (CC BY-NC-ND 4.0).

1. Introduction

With the advance of technology and computing a lot of communication moved from the classical paper into the digital format. Real estate marketing follows the same trend. Tsakiridou and Karanikolas (2019) citing a commercial report from 2017 of National Association of Realtors suggest that around 83% of the buyers identify their future property in online searches and only the rest use other methods. The COVID-19 pandemic, by imposing social distancing, accentuated this trend (Low *et al*., 2020).

Digital market, in general, and in real-estate in particular is made of 4 stake-holders (Low *et al*., 2020): the merchant (seller), the network, the publisher, and the customer. In this work we approach the problem from the publisher point of view which uses an Internet portal, where the seller or his representative lists the announcement advertising the property that is for sale or for rental. Such announcements contain explicit information about the property, but quite often (even mandatory on some portals) are accompanied by illustrative images. The content of these images is related usually to the view of the building or of the interior and respectively to the plans of the property.

The aim of the publisher is to offer the potential customer the best experience and thus to maximize the number of accesses and eventually income. Automatic analysis of images can be used into several directions as follows.

A first problem is the automatic identification of *cvasi-duplicates*. This is derived from a particular problem of the Romanian real-estate market. In US, in most cases, a realtor represents and commissions only the seller or only the buyer; in contrast in Romania, but also in other countries (Markoc and Fusun, 2021), many real-estate agents act as intermediary between the seller and the buyer and commission both the seller and the buyer. To maximize their interest, they often post fake announcements without real properties behind, or they duplicate an existing announcement, but with their contacts, trying to capture potential buyers and to insert themself in the transaction. The interest of the publisher (portal) is to remove fake competition and thus to remove such announcements. In the case of fake announcements, accompanying images are taken from other announcements, without or with small changes. We define a *cvasi-duplicate* as being an image that is identical or slightly changed when compared to one from an existing database.

A second problem is automatic labelling (auto-tagging) of images. Given an image, various information such as the type of the scene (outdoor view, indoor view, kitchen, bedroom), existing objects (useful especially for renting furnished/unfurnished properties) can be extracted and aggregated in the overall processing.

A third problem is to provide description such to enable searches to retrieve images similar as content. A user (potential buyer) interest of the

property can be measured by the amount of time spent on particular images. The user's experience can be improved if images with similar content are suggested.

The remainder of the paper is organized as follows: in the next section relevant real-estate image analysis solutions are briefed. Section 3 discusses the solution to cvasi-duplicates; section 4 discuses in joint manner classification and description for similarity. The paper ends with conclusions.

2. Prior Art

Once it was understood that images accompanying real-estate announcements contain useful data, multiple solutions to automatically extract information have been proposed. A review on the topic may be followed in (Koch *et al*., 2017). They summarize the images to be analysed as originating from 4 categories of sources (i) outdoor/street view, (ii) indoor images, (iii) floor plans, and (iv) aerial and satellite images. In Romania the vast majority are indoor images illustrating aspects of the property. Properties in development use plans and aerial images, as the actual site does not exist yet.

In this context of image availability, various applications have been developed. The usage of aerial/street images may be enlisted for 3D building reconstruction (Tascon-Vidarte, 2016), building style identification (Lee *et al*., 2015; Lee, 2013) or building elements (windows, arches, roof slant etc.) style recognition (Wenzel and Forstner, 2012).

More closely related to our problems, are the solutions developed on analysis of indoor images. Here, there is also a rather broad range of applications, that are typically divided into three (Koch, 2017): (i) the recognition of indoor scenes, (ii) the construction of 3D models, and (iii) the prediction of the market value of real estate. The construction of 3D models is outside the scope of this paper.

The solutions proposed for the recognition of indoor scenes originate in Torralba's *Places 365* work (Zhou *et al*., 2017) and several solutions adapted it to applications in real estate imaging. For instance, Nascimento (2018) used both global and local features to create a highly discriminatory representation of indoor scenes. Yet, in real life, an image is not only an indoor scene but also a plan, so a new approach is needed.

The prediction of the market value of the real estate can use additional information or only vision-based data. Recent approaches rely on a deep Convolutional Neural Network framework. Bessigner (2016) aggregated front-view images of houses with data such as house characteristics from the appraiser to evaluate the house price by using VGG16 and random forest. You (2017), using a hybrid model that combines visual descriptors from GoogleNet and long short-term memory (LSTM) recured time series to predict house prices. Liu (2018) fuses in a multi-instance model house images and attributes and follows with a ranking loss to estimate the prices.

Yet two observations may be made: to estimate the price one needs the image of the house and cannot do it for an apartment; secondly the most recognized solution is the one offered by the commercial company named Zillow (Poursaeed *et al*., 2018). While initially Zillow showed much success, more recently this solution failed to provide accurate estimates, forcing the company to withdraw it (Cao, 2021) after significant financial failure. Therefore, our purpose is more modest, aiming to simply offer support to customer.

3. Identification of Cvasi-duplicates

As said, we name a *cvasi-duplicate* an image that is too similar, as content, with one existing in the database. The two images (the query and the one found as being the source) are either identical, or obtained by a trivial image transform such as: horizontal flipping, crop (*e.g*. removing the source image logo), slight color palette change. From a computer vision point of view, this is an image retrieval problem.

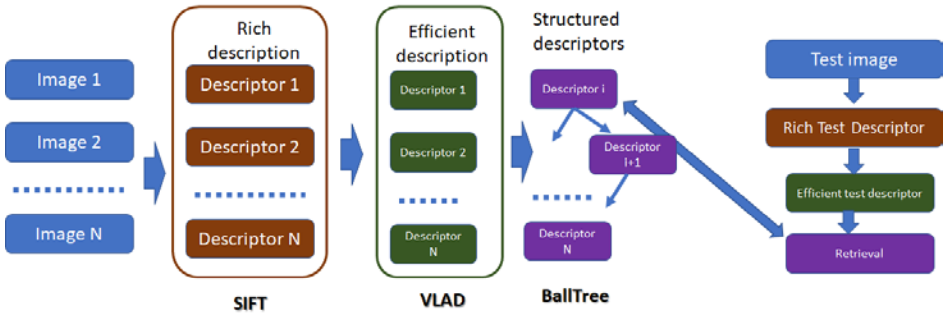


Fig. 1 – Cvasi-duplicate image retrieval scheme.

The solution used is inspired from (Arandjelovic, 2013). A schematic of the method may be followed in Fig. 1. As rich descriptor we used Scale Invariant Feature Transform with 128 length gradient vector, efficient description is based on VLAD (Very Large Aggregated Descriptor) with 64 clusters, the Ball Tree is used for efficient structure and search of new descriptor. For VLAD we have replaced the classical L2 distance with Hellinger Coefficient.

$$H(P,Q) = \frac{1}{\sqrt{2}} \sqrt{\sum_{i=1}^{k} (\sqrt{p_i} - \sqrt{q_i})^2} \tag{1}$$

where $\boldsymbol{P} = [p_1, p_2, \ldots p_k]$ and $\boldsymbol{Q} = [q_1, q_2, \ldots q_k]$ are the vectors to be compared

Various alternatives have been attempted and the performance was evaluated on a set of 250 query images taken out of a database of approximately 15000 images. Instead of SIFT descriptors, we tried ORB (Oriented FAST and Rotated BRIEF), but ORB fails systematically at horizontal flipping. SURF is another potential solution, but its use in practical applications is hindered by patent protection. The situation was similar for SIFT, but its main patent expired in 2020. Values for correct retrieval may be followed in Table 1. The used algorithm manages to return correctly cvasi-duplicates in 88.2% of the cases and in 100% if 5 images are returned.

Another set of tests refers to the robustness of the proposed algorithm to various transforms. In practice the most often met is cropping, such to remove the logo of original realtor, followed by horizontal flipping. We have tested other transforms/deformation and the results may be followed in Table 2. When a specific transform was studied all 250 were transformed. Adding noise was not met in practice, thus is less practically motivated; it is however motivated from a technical perspective as it may change the order of the gradient and affect SIFT descriptors.

Instead of the SIFT based descriptor, one may use CNN based, as are those described in further sections. Yet our experiments and prior art shows that CNN based descriptors are heavily sensitive to small changes based on slight noise. They offer a much better description when asking for images that have similar content, but they are fundamentally different as another scene is photographed.

Table 1

Correct retrieval rate when various alternatives of the algorithm have been tried. The database contained the 250 original images, plus 750 obtained by various small random transforms

Method	Original	ORB	SIFT length 256	SIFT length 32	VLAD -L2	VLAD – 128 clusters	VLAD – 32 clusters
Rate [%]	88.2	78	88.3	85.4	85.6	87.8	84.7

Table 2

Correct retrieval rate when various image deformations or transforms have been tried. Testing database contained 250 images added to a reference set of 15000

Transform	None	Horiz. Flip	Rotate up to 15 deg	Noise σ=1	Noise σ=2.5	Crop 90%	Palette change
Rate [%]	100%	100%	89.6	85.2	70.8	94.8	72.7

4. Auto-tagging and retrieval of real estate images

This problem was further divided into three problems: detection of objects of interest, global auto-tagging and image descriptor for similarity retrieval. All aspects have been approached using deep learning based frameworks.

Object detection and localization was carried using YoloV5 which is an incremental development from YoloV3 (Redmon, 2018). The solution was trained on a dataset with 25 classes. The classes, number of cases and the reached performance may be followed in Fig. 3. Yolo reports bounding boxes and a detection is "true" if Intersection Over Union (IOU) with ground truth is larger than 0.5. A visualization of detection examples can be seen in Fig. 2. To eliminate the overlapping detection, a layer of non-maxima suppression has been implemented. Objects detected are further used to validate information related to "furnished", which is especially useful for rented properties announcements.

Global auto-tagging assumes given an image as an input and returning a label. We have constructed a database with approximately 70.000 images (from public Places365 database and our own set of real-estate properties) which has been used for training and tested on our own annotated set of approximately 10000 images. The number of classes used in the database was 27 and chosen as being relevant either as appearance in announcement either in impacting the price of the property: *'map', 'rendering', 'plan', 'indoor - empty', 'kitchen', 'living-room', 'dining-room', 'bathroom', 'childroom','bedroom','utility-room','indoor - balcony' 'indoor - staircase', 'outdoor - kennel', 'outdoor-façade - building', 'outdoor-facade - office', 'outdoor-facade - house', 'outdoor-parking', 'outdoor-yard', 'outdoor - porch/deck', 'outdoor - garage', 'outdoor - construction-site', 'outdoor - terrain', 'indoor - lobby','outdoor - entrance','swimming-pool - indoor', 'swimming-pool - outdoor', 'indoor - office'*. Various architectures and training strategies have been tested. All models have been pretrained on the full Places365 database. Results may be followed in Table 3.

However, one may go further. Another problem that needs to be addressed is that further processing of the announcement, given a query, one needs to identify similar properties. While announcement processing is outside the scope of this paper, the problem of retrieving images that are similar from a content point of view stands in scope. A trained deep convolutional network offers by activation values in various layers, image descriptors that can be used for retrieving images that are similar in content. A typical solution to this problem is to ask the network to build simultaneously good class separation and discriminative embeddings; those can be obtained by using a triplet loss strategy (Gordo and Larlus, 2017; Heaven, 2019). Yet we prefer to use a large margin (Florea *et al.*, 2020; Khin *et al*., 2016) loss function, which is defined as:

$$\mathcal{L}_{\mathcal{M}} = \sum_{i=1}^{N} \left(\left\| \frac{\mathbf{e}_i}{\|\mathbf{e}_i\|_2} - \frac{\mathbf{c}^j}{\|\mathbf{c}^j\|_2} \right\|_2 - \frac{1}{C-1} \sum_{k=1,k\neq j}^{C} \left\| \frac{\mathbf{e}_i}{\|\mathbf{e}_i\|_2} - \frac{\mathbf{c}^k}{\|\mathbf{c}^k\|_2} \right\|_2 \right) \quad (2)$$

where $\boldsymbol{e}_i$ describes an embedding associated with a data labelled as being from class i ($y_i = j$), $\boldsymbol{c}_j$ is the centroid of the same class j out of C and N is the number of instances. The total loss function is, then given as:

$$\mathcal{L}_{\mathcal{T}} = \lambda_1 \mathcal{L}_{\mathcal{S}} + \lambda_2 \mathcal{L}_{\mathcal{M}} \quad (3)$$

where $\mathcal{L}_S$ is the standard cross entropy while λ_1 and λ_2 are weighting constants. Using the large margin concept and taking $\boldsymbol{e}_i$ embeddings from the last layer before the decisional one, we obtain the description required for retrieval of images that are similar in content. Furthermore, we also have a very large corpus of unlabelled data taken from our real-estate collection. Using the Margin-Mix semi-supervised strategy (Florea, 2020) we are able to further improve the performance as listed in the lower part of Table 3.



Fig. 2 – Input image and auto-tagging. Global is correct (“Kitchen”) and bounding boxes found by Yolo do miss the table and suffer from lack of non-maxima suppression.

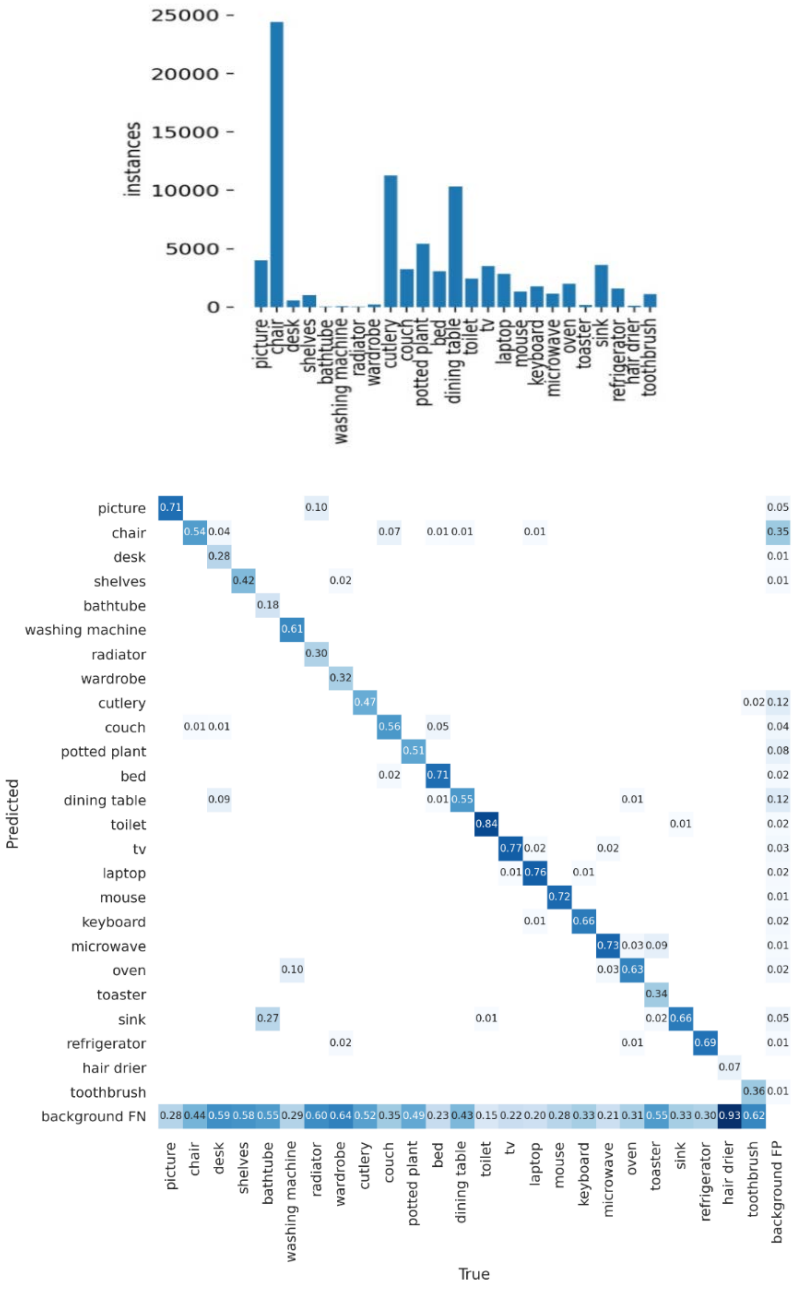


Fig. 3 – Object localization in real-estate images. The solution, Yolov5 has MobilenetV2 as backbone. Top plot: number of instances existing in the dataset. Bottom plot: confusion matrix of the final solution.

Table 3
Performance (accuracy) of the global auto-tagging system

Trained system			Performance
Architecture	**Optimization**	**Learning**	**[%]**
Resnet 18	ADAM	Supervised	66.53
Resnet 18	SGD	Supervised	71.40
Resnet 18	ADAM	Supervised	67.25
Resnet 50	SGD	Supervised	71.20
Resnet 50	SGD	Supervised	72.90
Resnet 18	SGD	Margin-Mix	73.50
Resnet 50	SGD	Margin-Mix	75.50

Global image labels are further used in real-estate announcement classification and organization. Embeddings, corroborated with other data listed in an announcement (*e.g*. location, price, surface) are used to retrieve similar listings, which are all brought on a single screen such that the time spent by a potential buyer searching for the right house is more efficiently used.

4.1. Implementation details

The auto-tagging part that is built using deep learning tools uses PyTorch framework. The training was carried on a Lenovo Yoga with MTX Nvidia GPU with 8GB of Memory. A training takes about 12 hours. In testing, the application uses a more powerful station having RTX 360 and labels a batch of 10 images (the averages of images per announcement) in less than 0.2 secs.

If less resources are available, instead of the more memory hungry ResNet 50, a ResNet -18 may be used. Tests on a laptop with Nvidia 950, 2GB GPU keep the duration in the same range.

5. Conclusions

In this paper we showed how state of the art computer vision algorithms and solutions can be used to address practical problems related to real-estate imaging. The application is slanted to be integrated together with other functionalities so to fully have an “intelligent” real-estate portal.

Acknowledgements. The work has been funded by the project “Innovative Artificial Intelligence systems in the field of real estate portals” (Sisteme inovative de Inteligență Artificială în domeniul portalurilor imobiliare), online number 137-221-A2, MySMIS number: 129132. .

REFERENCES

Cao S., *Zillow's Fixer-Upper Disaster: Why Zestimate Couldn't Get Home Pricing Right*, Observer, 2021.

Florea C., Badea M., Florea L., Racoviteanu A., Vertan C., *Margin-Mix: Semi-Supervised Learning for Face Expression Recognition*, In European Conference on Computer Vision (pp. 1-17), Springer, Cham. 2020, August.

Gordo A., Larlus D., *Beyond instance-level image retrieval: Leveraging captions to learn a global visual representation for semantic retrieval*, in CVPR, pp. 6589-6598, 2017.

Heaven Douglas, *Why deep-learning AIs are so easy to fool*, Nature 574 (7777): 163-166 (October 2019).

Khin A.A., Chau W.H., Chuan C.S., Tong L.K., *Factors Affecting for Online Marketing Strategy of Business Development in Malaysia*. J. Econ. Manag. Trade 2016, 11, 1-17 (2016).

Koch D., Despotovic M., Leiber S., Sakeena M., Döller M., Zeppelzauer M., *Real estate image analysis: a literature review*, Journal of Real Estate Literature, 27(2), pp. 269-300 (2019).

Lee S., Maisonneuve N., Crandall D.J., Efros A.A., Sivic J., *Linking past to present: Discovering style in two centuries of architecture*, In 2015 IEEE International Conference on Computational Photography (ICCP), pp. 1-10 (2015).

Lee D.-H., *Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks*, in International Conference on Machine Learning Workshops, 2013.

Liu X., Xu Q., Yang J., Thalman J., Yan S., Luo J., *Learning multi-instance deep ranking and regression network for visual house appraisal*, IEEE Trans Knowl Data En 30(8):1496-1506 (2018).

Low S., Ullah F., Shirowzhan S., Sepasgozar S.M., Lin Lee C., *Smart digital marketing capabilities for sustainable property development: A case of Malaysia*, Sustainability, 12(13), p.5402 (2020).

Markoc Ilkim, Fusun Cizmeci, *Ethics in real estate: agency practices in Istanbul*, International Journal of Housing Markets and Analysis (2021).

Nascimento G., Laranjeira C., Braz V., Lacerda A., Nascimento E., *A robust indoor scene recognition method based on sparse representation*. In M. Mendoza and S. Velastn (Eds.), Progress in Pattern Recognition, Image 31 Analysis, Computer Vision, and Applications. CIARP 2017. Lecture Notes in Computer Science, Volume 10657, Cham, pp. 408{415. Springer (2018).

Poursaeed O., Matera T., Belongie S., *Vision-based real estate price estimation*. Machine Vision and Applications, 29(4), pp.667-676 (2018).

Redmon J., Farhadi A., *Yolov3: An incremental improvement*, arXiv preprint arXiv:1804.02767 (2018).

Tascon-Vidarte J.D., *Floor plans from 3d reconstruction of indoor environments*, In 2016 XXI Symposium on Signal Processing, Images and Artificial Vision (STSIVA), pp. 1-7 (2016, Aug).

Tsakiridou Eleni, Nicolas Karanikolas, *Real estate in the web and social media era, the Greek reality*, RELAND: International Journal of Real Estate & Land Planning 2: 28-43 (2019).

Wenzel S., Forstner W., *Learning a compositional representation for facade object categorization*, ISPRS Annals of Photogrammetry, Remote Sensing and Spatial Information Sciences I-3, 197-202 (2012).

You Q., Pang R., Cao L., Luo J., *Image-based appraisal of real estate properties*, IEEE Trans Multimed 19(12): 2751-2759 (2017).

Zhou B., Lapedriza A., Khosla A., Oliva A., Torralba A., *Places: A 10 million Image Database for Scene Recognition*, IEEE Transactions on Pattern Analysis and Machine Intelligence, 40(6), pp.1452-1464, 2017.

ANALIZA AUTOMATĂ A IMAGINIILOR CU CONȚINUT IMOBILIAR PENTRU APLICAȚII DE REGĂSIRE ȘI CLASIFICARE

(Rezumat)

În prezent, sectorul imobiliar beneficiază și el de progresul tehnologiei. Imaginile digitale însoțesc publicitatea pentru închirierea sau vânzarea proprietăților. Analiza imaginilor și inteligența artificială contribuie la înțelegerea acestor imagini către o experiență mai bună a consumatorului. În această lucrare arătăm cum pot fi utilizate eficient regăsirea prin conținut și descrierea imaginilor pentru a izola cvasi-duplicatele și pentru a identifica exemple similare cu cele care au atras interesul privitorului. Sunt identificate trei probleme legate de imagistica imobiliară și sunt propuse soluții care sunt evaluate pe baze de date mari.