Saliency-Based Keypoint Reduction for Augmented-Reality Applications in Smart Cities

Simone Buoncompagni[1(✉)], Dario Maio[1], Davide Maltoni[1], and Serena Papi[2]

[1] DISI, Università di Bologna, Mura Anteo Zamboni 7 40126, Bologna, Italy
simone.buoncompagni2@unibo.it
[2] CIRI ICT, Università di Bologna, via Rasi e Spinelli 146 47521, Cesena, Italy

Abstract. In this paper we show that Saliency-based keypoint selection makes natural landmark detection and object recognition quite effective and efficient, thus enabling augmented reality techniques in a plethora of applications in smart city contexts. As a case study we address a tour of a museum where a modern smart device like a tablet or smartphone can be used to recognize paintings, retrieve their pose and graphically overlay useful information.

Keywords: Saliency-based ranking · Keypoint local descriptors · Smart city · Augmented reality

1 Introduction

The growth of mobile devices equipped with high quality displays, high resolution cameras and high processing capabilities allows new computer vision applications to be deployed. In particular in the context of smart cities, augmented reality is an enabling technology for a number of applications in tourism, arts and intelligent buildings since as defined in the European context the presence of cultural facilities is a key indicator for smart cities quality evaluation [12].

It is well-known that accurate object recognition and pose detection are key building blocks to develop effective Augmented Reality (AR) applications. Moreover, when artificial landmarks (e.g., 2D-bar codes, beacons, etc. [5] [6]) cannot be used, recognizing objects and retrieving their pose in real-time can be very challenging, especially on resource-constrained platforms such as mobile devices. In [13][14] different strategies have been proposed in order to reduce the number of keypoints (and corresponding local descriptors) that need to be matched. In [1] we recently introduced a pose detection approach founded on a Saliency-based keypoint selection and reduction that has been proved to be very effective for the problem at hand.

In this work we extend approach [1] by including an object recognition phase to be carried out before pose estimation, and we design a specific application to perform a museum tour with the aid of AR. The paper is organized as follows: in section 2 we present an overview of the Saliency-based keypoint selection method introduced in [1]; in section 3 we extend our previous approach in the context of the tour of a museum; finally, in section 4 we draw some conclusions.

V. Murino et al. (Eds.): ICIAP 2015 Workshops, LNCS 9281, pp. 209–217, 2015.
DOI: 10.1007/978-3-319-23222-5_26

2 Saliency-Based Keypoint Selection: An Overview

In this section we summarize the method we proposed in [1] for Saliency-based keypoints selection.

Given an object, a preliminary training step is performed to define the object model based on its most salient keypoint descriptors. The training set is composed by a single reference image I^{ref} of the object acquired in neutral viewpoint and lighting conditions and by a set of *N* generated images I^1, I^2, ..., I^N which depict the same object under different conditions. A generic transformed image $I^l = Transf_l(I^{ref})$ is obtained by applying a transformation (e.g. 2D homography, a 3D projection, a light changing function) to the reference image.

To evaluate saliency, keypoints detection on the reference image I^{ref} is firstly performed and then each keypoint is mapped on the transformed images by applying $Transf_l$ functions. Descriptors for all keypoints are computed and a global analysis is performed to rank the keypoints by saliency and to retain only the *m*-best ones to characterize the object model. Highly salient keypoints are excellent candidates for the matching since focusing on them not only reduces the computational load but also improves keypoint matching accuracy.

Even if our approach is independent of the keypoint detector and local description, to maximize efficiency in [1] we focused on FAST detector and BRIEF descriptors. Furthermore, we proved that working in the Opponent color space [3] (instead of the RGB space) increases the robustness with respect to light changes.

For a given object, let $\boldsymbol{x}_i = (u_i, v_i) \in I^{ref}$ be a keypoint selected by the FAST detection algorithm [11] and be s_i its strength returned by the FAST algorithm itself. The set of all keypoints of the reference image is $K_d(I^{ref}) = \{(\boldsymbol{x}_i, s_i): \boldsymbol{x}_i \in I^{ref}, i = 1, \dots, J\}$. For each $\boldsymbol{x}_i \in K_d(I^{ref})$, we define with $descr: (\Re^2, \Re^S \times \Re^S) \rightarrow \Re^L$ the function that computes BRIEF descriptor [2] for a keypoint $\boldsymbol{x}_i$ according to the image patch $P(\boldsymbol{x}_i)$ of size S×S centered on $\boldsymbol{x}_i$. Given the nature of BRIEF, $\mathbf{b}_i = descr(\boldsymbol{x}_i, P(\boldsymbol{x}_i)))$ is a binary vector. Therefore, two binary vectors $\mathbf{b}_i$ and $\mathbf{b}_j$ are compared by using the Hamming distance $H(\mathbf{b}_i, \mathbf{b}_j)$ that can be computed very efficiently through a bitwise XOR operation followed by a bit count.

Keypoint saliency is expressed in terms of detectability, distinctiveness and repeatability (see [1] for equations), defined as follows:

- The *distinctiveness* $D(\boldsymbol{x}_i)$ of a keypoint $\boldsymbol{x}_i \in K_d(I^{ref})$ is proportional to the diversity among the $\boldsymbol{x}_i$ descriptor and the descriptors of other keypoints $\boldsymbol{x}_j \in K_d(I^{ref})$, $j \neq i$ in the same image.
- The *repeatability* $R(\boldsymbol{x}_i)$ of a keypoint $\boldsymbol{x}_i \in K_d(I^{ref})$ is proportional to the similarity among its descriptor $\boldsymbol{b}_i$ and the descriptors of corresponding keypoints under a set *T* of given transformations (e.g. viewpoint and lighting).
- The *detectability* $F(\boldsymbol{x}_i)$ of a keypoint depends of the score values returned by the keypoint detection algorithm (i.e. in FAST, the score is the corner strength [2]) and quantifies the aptitude of a given keypoint to be detected under various viewpoint and lighting changes. The detectability of a keypoint $\boldsymbol{x}_i \in K_d(I^{ref})$ is simply an average (normalized in the range [0,1]) over the scores of all keypoints in the original image and its transformed versions.

It is worth noting that while detectability is related to keypoint stability under transformation, repeatability and distinctiveness are related to the discriminant power of descriptors. Detectability, distinctiveness and repeatability are finally combined in order to determine the *keypoint Saliency S*, as follows:

$$S(\boldsymbol{x}_i) = \omega_R R(\boldsymbol{x}_i) + \omega_D D(\boldsymbol{x}_i) + \omega_F F(\boldsymbol{x}_i) \quad (1)$$

where ω_R, ω_D and ω_F are weights assigned to repeatability, distinctiveness and detectability, respectively.

3 A Case Study: A Museum Tour with Augmented Reality

The Saliency-based approach proposed in [1] is here extended with a (pre)matching phase and applied to the painting recognition and pose estimation, which constitute useful building blocks to develop AR based museum tour. A number of AR solutions in the field of cultural heritage and mobile multimedia guides have been recently proposed [7][8][9][10]. Authors of [7] and [9] introduced an exhaustive overview of the main challenges related to conception, implementation, testing and assessment of a smart museum. In PALM-Cities Project [10] technologies such as NFC and QR Codes have been adopted to handle the interaction with the user whereas in [7] an hybrid approach based on markerless tracking plus a rotation sensor is used to allow free movements of the user mobile device.

Similarly to [7], in our application the user is expected to enjoy paintings in a markerless environment by interacting with a mobile device (e.g. tablet, smartphone or smart glasses) provided with a camera that captures videos of paintings under different conditions (i.e., moderate changes of viewpoint and lighting).

Once a painting has been recognized and its pose has been retrieved the application can properly superimpose to the live camera view useful pictorial or textual information concerning the painting itself (see Fig. 1 for an example).

An overview of the approach is presented in Fig. 2: during the training phase we use a single reference image of each painting to compute the painting model including only the most salient keypoints. Models are then stored in a database which is made available to the user's mobile device.

In this study we consider 10 famous paintings (see Fig. 3). For each painting p we downloaded the reference image I_p^{ref} from the web and printed it on paper (A3 format). Paintings were then hanged to the walls of our lab to simulate a museum room.

Fig. 1. Example of augmented information geometrically coherent with the painting pose.

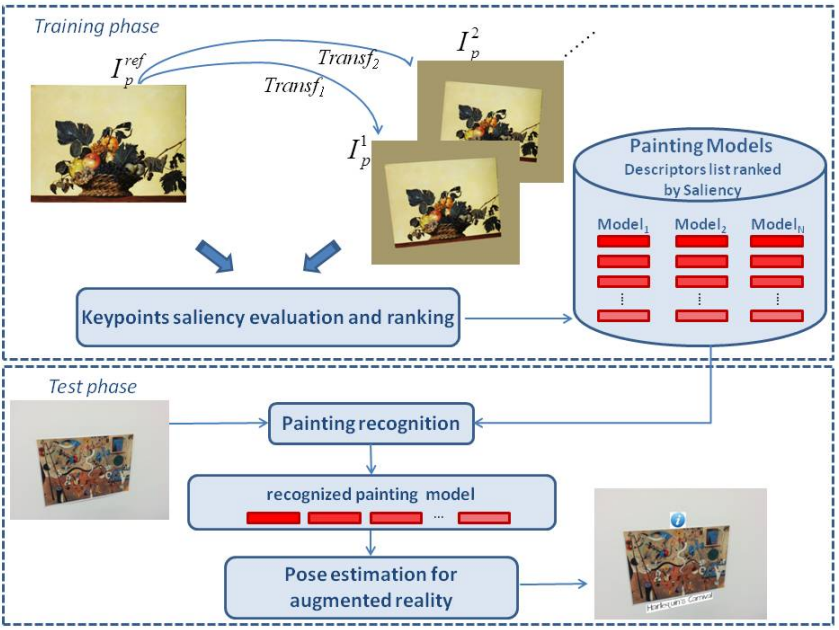


Fig. 2. Overview of the proposed application based on keypoint saliency evaluation.

Fig. 3. The 10 famous paintings we consider in our study.

For each reference image, we generated 80 artificial transformations to be used for the training phase: the variations considered are random homographic transformations within predefined parameter ranges.

Test was performed using a smart device and capturing videos of each printed painting while moving in front of the painting; for each video we selected 30 frames characterized by different lighting and pose conditions, hence our test set is composed of 300 images (see Fig. 4 for some examples).

We performed two different experiments: the former to evaluate recognition accuracy and the latter to evaluate the correctness of the estimated pose. In both these experiments our Saliency-based ranking is compared to a standard FAST score-based ranking. For each test image I^{test}, the painting recognition phase is implemented as follows:

- all keypoints are extracted (FAST) and their local descriptors (BRIEF) computed;
- for each painting model I_p^{ref}, characterized by its m-most salient keypoints $K_d^m(I_p^{ref})$:
  - we associate each keypoint in $K_d^m(I_p^{ref})$ to the keypoint in I^{test} with smallest Hamming distance (between BREIF descriptors);
  - we enforce geometrical constraints among keypoint correspondences using RANSAC algorithm [4] to filter out outliers;
  - the set of inliers returned by RANSAC is then used to compute a similarity score Φ between I^{test} and I_p^{ref} as follows:

$$\Phi(I^{test}, I_p^{ref}) = \frac{\#\ Ransac\ Inliers}{\#\ K_d^m(I_p^{ref})} \tag{2}$$

- finally, recognition is performed according to maximum similarity.

Fig. 4. Samples of the dataset used in our case study: (a) and (d) are the reference images of two different paintings, (b), (c), (e), (f) are test frames acquired live with a tablet.

Tests have been repeated for n = 300 and for different values of m ranging from 1% to 100% of the total number of keypoints. This allows to evaluate the effect on recognition of progressive reduction of the keypoint number.

In Table 1 we show the recognition rate obtained by considering only the most salient descriptors when our Saliency-based ranking and a standard FAST-scores ranking are applied. In general we can observe that our method is more effective than the FAST-scores based one. Moreover it turns out that by applying our Saliency evaluation a lower percentage of keypoints is sufficient to reach top performance with respect to FAST-scores ranking (only 4% of the keypoints for our approach versus 15% for FAST-scores).

Table 1. Recognition results for Saliency-based selection and Fast-scores selection.

***m*-most salient considered keypoint (%)**	**"Saliency-Based" Recognition rate (%)**	**"Fast-based" Recognition rate(%)**
1	74.58333	53.75
2	92.5	80.41667
3	97.91667	90.83333
4	**98.75**	90
5	98.75	94.16667
10	97.91667	95.83333
15	96.66667	**97.91667**
20	97.08333	97.91667
100	88.75	88.75

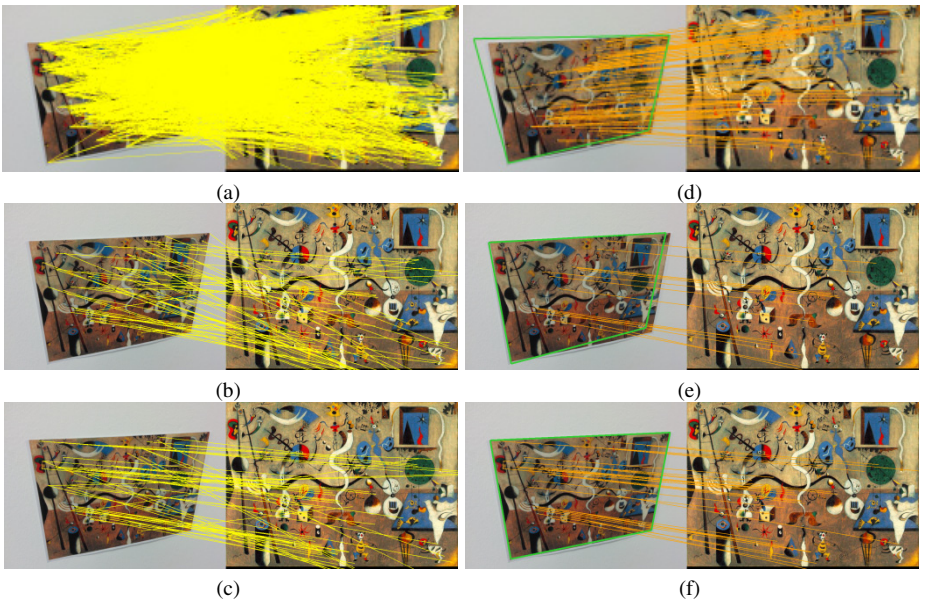


Fig. 5. Painting transformation recovery through RANSAC algorithm by taking as input: (d) all FAST keypoints; (e) 5% m-best keypoints ranked according to FAST score, (f) 5% m-best keypoints ranked according to our Saliency-based approach. Yellow segments (a), (b) and (c) denote initial keypoint pairing and orange segments (d), (e) and (f) final RANSAC inliers; the green rectangle denotes the homographic transformation inferred by RANSAC.

A second evaluation has been carried out to assess correctness and computational load of pose estimation. Since a painting can be considered as a full planar object, pose is computed by estimating a homographic transformation through the RANSAC algorithm [4] given a set of keypoint correspondences with the reference model.

Fig. 5 shows the result of pose estimation for a painting sample by considering three different cases: (a) model including all keypoints, (b) only 5% *m*-best keypoints selected according to FAST scores and (c) only 5% *m*-best keypoints selected according to our approach.

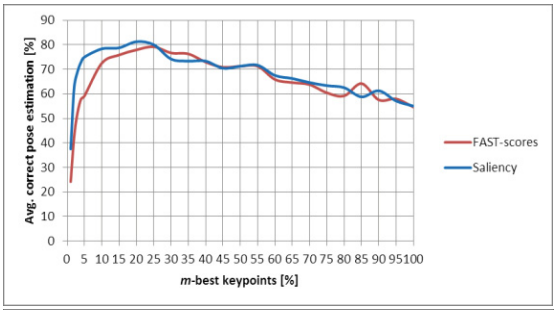
Although RANSAC is somewhat robust with respect to outliers, the advantages of using only relevant keypoints are here evident in terms of precision of the recovered viewpoint transformation. We also note how our ranking leads to consolidate a higher number of inliers and therefore a better viewpoint estimation with respect to the FAST-scores based selection.

To numerically quantify pose estimation accuracy we manually marked (as ground truth) the four painting corners both for each reference image and each test frame. A pose is then considered correct when the projected painting corners (according to the estimated homography) have a spatial distance from the corresponding ground truth lower than a prefixed threshold.

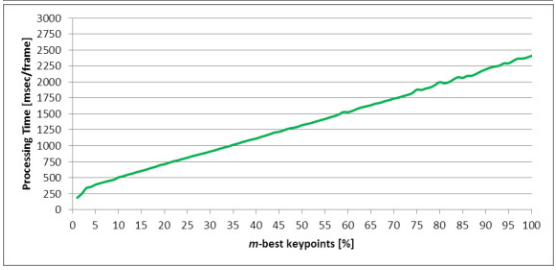
In the graph of Fig. 6a we show the percentage of "correct pose" estimation averaged over all 300 tests images.

We can easily note that, in both cases, the curves have an increasing trend up to a relatively small value *m* of best keypoints and then start decreasing. The optimal percentage of keypoint falls in the range [5%, 20%] in our approach, and in [15%, 30%] when selection is performed according to FAST scores.

Fig. 6b shows the average processing time for a single frame analysis as function of the keypoints percentage. For this experiment we used a Samsung ATIV Smart PC (Intel Atom Processor Z2760 1.5 Ghz) tablet device. Even if the code (written in C# for .NET) was not highly optimized, by selecting a percentage of keypoints below 5% we can provide a frame rate from 3 to 5 frame/s, ensuring, at the same time, good accuracy in terms of object recognition and pose estimation.



(a)



(b)

Fig. 6. (a) Average percentage of correct poses by varying the percentage of the keypoints ranked both according to Fast-scores and our Saliency-based approach; (b) Average processing time (milliseconds) required for a single frame analysis including painting recognition and pose estimation on Intel Atom Processor Z2760 1.5 Ghz.

4 Conclusions

In this paper we proved the feasibility of a markerless AR application running on mobile devices. Initial results with a limited database (10 paintings, 300 test poses) are quite promising. The main strength of the proposed Saliency-based ranking and selection relies in the significant reduction of the amount of features to be matched, thus allowing real-time implementation on resource-limited computer architectures without compromising recognition accuracy. In the future, we intend to study the efficacy and efficiency of this technique when scaling to larger datasets and with different combinations of feature detectors / local descriptors. On the one side, we expect that painting recognition accuracy and efficiency can be negatively affected when a large number of painting models are stored in the database. However, for the application considered this is not a serious problem since a single tag (e.g. NFC, e-beacon, etc.) could be placed inside each room to coarse localize the user and restrict the database search to the paintings located inside the current room.

References

1. Buoncompagni, S., Maio, D., Maltoni, D., Papi, S.: Saliency-based keypoint selection for fast object detection and matching. Pattern Recognition Letters **62,** 32–40 (2015)
2. Calonder, M., Lepetit, V., Strecha, C., Fua, P.: BRIEF: binary robust independent elementary features. In: Daniilidis, K., Maragos, P., Paragios, N. (eds.) ECCV 2010, Part IV. LNCS, vol. 6314, pp. 778–792. Springer, Heidelberg (2010)
3. Van De Sande, K.E., Gevers, T., Snoek, C.G.: Evaluating color descriptors for object and scene recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence **32**(9), 1582–1596 (2010)
4. Fischler, M.A., Bolles, R.C.: Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. Communications of the ACM **24**(6), 381–395 (1981)
5. Parikh, D., Jancke, G.: Localization and segmentation of a 2D high capacity color barcode. In: IEEE Workshop on Applications of Computer Vision, WACV 2008, pp. 1–6, January 7–9, 2008
6. Furht, B.: Handbook of augmented reality, vol. 71. Springer, New York (2011)
7. Miyashita, T., et al.: An augmented reality museum guide. In: Proceedings of the 7th IEEE/ACM International Symposium on Mixed and Augmented Reality. IEEE Computer Society (2008)
8. Damala, A., Marchal, I., Houlier, P.: Merging augmented reality based features in mobile multimedia museum guides. In: CIPA Conference on Anticipating the Future of the Cultural Past, 2007, October 1–6, 2007
9. Damala, A., et al.: Bridging the gap between the digital and the physical: design and evaluation of a mobile augmented reality guide for the museum visit. In: Proceedings of the 3rd International Conference on Digital Interactive Media in Entertainment and Arts. ACM (2008)
10. Caridi, A., Coccoli, M., Volpi, V.: Wolfsoniana smart museum. a pilot plant installation of the PALM-cities project. In: UMAP Workshops (2013)

11. Rosten, E., Porter, R., Drummond, T.: Faster and better: A machine learning approach to corner detection. IEEE Transactions on Pattern Analysis and Machine Intelligence **32**(1), 105–119 (2010)
12. http://www.smart-cities.eu/download/smart_cities_final_report.pdf
13. Carneiro, G., Jepson, A.D.: The quantitative characterization of the distinctiveness and robustness of local image descriptors. Image and Vision Computing **27**(8), 1143–1156 (2009)
14. Hartmann, W., Havlena, M., Schindler, K.: Predicting matchability. In: Conference on Computer Vision and Pattern Recognition (2014)