Acta Universitatis Sapientiae
Electrical and Mechanical Engineering, 3 (2011) 15-25

Photo Tagging in Urban Environments

Rezső HANDBAUER, Csaba SIMON, Markosz MALIOSZ

Department of Telecommunications and Media Informatics,
Budapest University of Technology and Economics, Budapest, Hungary
e-mail: {handbauer; simon; maliosz}@tmit.bme.hu

Manuscript received June 10, 2011; revised December 10, 2011.

Abstract: In this paper we investigate the applicability of the photo tagging to geo location in urban environments. We rely on photos taken at the location to be identified and not on the geographical coordinates. The difficulty of the proposal was to identify the building/landmark based on the photo provided on-line by the user. Our goal was to provide a working solution with a reasonably fast reaction time for urban environments. We have shown that the combination of division of the image and the color-based comparison with the original SIFT algorithm significantly improves the comparison process.

Keywords: Photo tagging, augmented reality, image comparison.

1. Introduction

Linking real and virtual worlds is vastly researched and experimented by the research community. The proposed solutions are based on the idea of providing a solution that maps the virtual world on the real one, extending the elements of the real world with useful information and/or properties. Several proposals targeted the urban environments, when different locations were associated with meta-information [1], [2], allowing the civil groups to interact and change the perception of others on the respective locations. By now we can state that the community agrees that there is a need for such solutions, the current research being focused on the proper technological solutions that satisfy the requirements of the various target applications.

Most of the solutions rely on *tag*s associated with real world locations and meta-information is indexed by these tags. The proposed solutions chiefly differ in the tagging system and the way these tags are obtained by real world characters. This research area is strongly related to the topic of location based services offered to mobile users [3]. Although GPS (Global Positioning System) is available in many devices, it is not always available in urban and indoor environments [4]. A different idea was to use mobile cell information to locate the device, but operators are reluctant to offer such data and the precision of this solution is not high enough [5]. Mobile tagging optimizes the barcodes to mobile environments, but it requires the dispatch of the tags on buildings and outdoor locations [6].

In this paper we investigate the applicability of the photo tagging to geo location in urban environments. We rely on photos taken at the location to be identified instead of geographical coordinates. The advantage of this solution is that the buildings or landmarks are already in place, and the loss of signal does not affect its operation. The difficulty of the proposal was to identify the building/landmark based on the photo provided on-line by the user. There were several proposals that deal with such problems, but the solutions are not public and have been publicized only through demonstration events. Our goal was to provide a working solution with a reasonably fast reaction time for urban environments. The use case of the proposed photo tagging solution is that an integrated system should be able to offer extra information on urban locations based on pictures taken and uploaded real time with smartphones.

In the next section we present the image processing issues relevant to our topic, and then we present the proposed solution. In section 4 we analyze the performance of our proposal and finally we conclude our paper.

2. Image processing aspects

As explained in the previous section, our proposal requires the image-based identification of buildings and landmarks. There have been proposed several solutions, but each of them had problems during operation. The most advanced and successful image comparison solutions have been developed in face recognition [7].

Compared to that area of image comparison, our case has several particularities. The pictures sent in by the users most probably will not be taken exactly from the same position as the reference ones, therefore the comparison of these types of images have some specific properties. The most important aspects that make the image comparison harder are the following ones.

- The pictures are not taken from the same angle.
- The pictures are taken from different distances from the building.

- There might be several distracting details on the picture taken by the user.
- Changing light conditions, depending on the time of the day, weather, etc.
- The quality of the picture taken by the user probably is different (typically worse).

In the urban environment most of the landmarks are buildings, which have distinctive edges. Research in computer imaging has widely studied the issue of edge detection [8], this one being the starting point of most solutions. Several interest points of an image can later be extracted from edges and these interest points are later used as inputs by other image comparison algorithms.

The most used technique in edge detection is based on the intensity gradient of the image. Canny's algorithm is still the most used one and is based on five, relatively simple, easy-to-implement steps [9], [10].

Edges are characteristic to a given image, but due to the issues enumerated above we need a much more robust solution. Several solutions have been proposed, which operate on a larger set of features, called keypoints. These solutions handle image translation, scaling, rotation, local geometric distortion and minor changes in illumination or color. The most known among these solutions is the Scale Invariant Feature Transform (SIFT) [11], and its enhancement, the Gradient Location-Orientation Histogram (GLOH) [12]. Relatively recent proposal is the Speeded Up Robust Feature (SURF) [13], partly inspired by the SIFT descriptor. SURF is faster than SIFT and is more robust to image transformation and noise. Nevertheless, the SIFT algorithm is better supported by programming libraries and for this reason we used it in our solution. As explained in the following section, we opted for a modular solution and therefore it can be replaced with newer/better algorithms.

The generic image recognition systems perform well if the objects on the images have been pictured under the same angle. Large angle-deviations result in false positives or no matches. In order to reduce the occurrence of such problems we pre-process the images. Using the algorithms presented in this section we proposed a framework that is able to recognize pictures, enabling a photo tagging application in urban environments.

3. The proposed image processing framework

Our solution is a combination of image processing mechanisms. We tried to speed up the comparison process by removing the irrelevant parts of the image or to use new information in order to exclude false positives.

A. Image pre-processing

The image pre-processing transforms the image into a canonic form. We expected that this step would improve the process by increasing its accuracy. As it will be presented later in section 4 this came with the cost of increased duration, therefore the feasibility of this step should be further studied.

We apply the Canny algorithm to detect the edges of the image. Even if this is an old solution, it is accurate, it has several implementations, its implementation is relatively simple and it is fast [14]. Then we apply a Hough transform [15] to the resulting image to extract the straight edges and lines from the image. In most of the cases in urban environment this yields the parallel edges of a building. An illustrative example of the above steps is presented in *Fig. 1*.

Figure 1: The picture a.) before any transformation, b.) after edge detection, c.) after Hough transform.

Finally we transform the image, which means the rotation of the object (building/landmark) on the picture. The reference points needed for this rotation are those obtained through the Hough transform, but the AffineTransform [16] process is applied to the original color image. Therefore we keep the extra information of the original image that was lost through the first two steps. An illustrative example of this transform is depicted in *Fig. 2.*, where a boat is rotated so that the top edge of the cabin becomes horizontal.

Figure 2: The picture a.) before rotation, b.) after rotation.

Note that these steps are useful only if we use a SIFT (or similar) algorithm to compare the images. E.g., if we use color based comparison, then the fill color would greatly bias the process.

B. Keypoint based comparison

The image comparison is primarily based on the SIFT algorithm, as already mentioned.

This process can be greatly improved if we have location information and the estimate of its accuracy. E.g., we might have a coordinate and then we can use the accuracy of this location information as the radius of a circle that probably contains the real location of the user. We can restrict the search, because in the case of a precision of 50 m probably we have 20-50 buildings or landmarks to run the search on.

We should compare our image against the images of the buildings closer to the coordinate: if the location information was accurate, then we do not waste time. Otherwise we still can increase the investigated area.

C. Color based identification

Our color based identification is built with the help of the Java Advanced Imaging (JAI) API [17]. This library allows us to get the color data of an image in several points of it, then builds a matrix that represents the image itself. Afterwards each image is represented by a matrix with the dimensions of 25×3. Based on our experiments the aggregation of the 15×15 pixel region of the image into one matrix element gives satisfactory results. Even if we use larger regions, the accuracy of the mechanism does not improve significantly, on the other hand its runtime is drastically increased. The aggregation of the pixel information is as follows: the RGB values of each pixel within a region are summed up, then the result is divided with the total number of pixels.

The size of the image does not influence the speed and quality of the process, because we resize every image to 300×200 pixel^2. Due to this and because the algorithm is much simpler than the SIFT, the comparison is much faster.

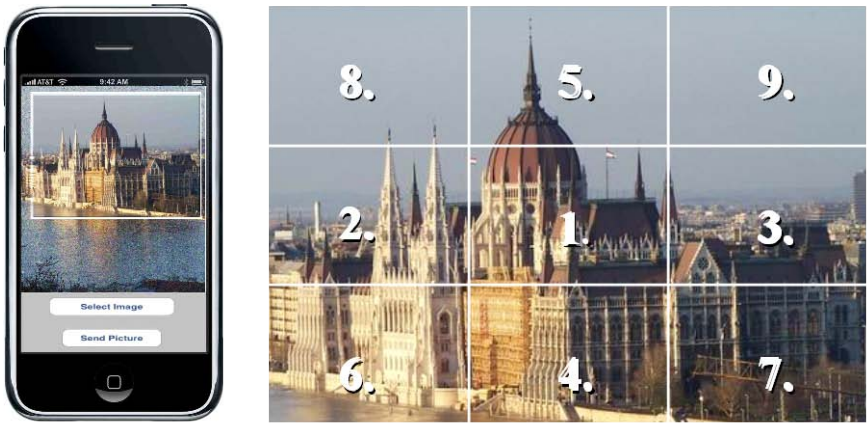
D. Speeding up the image comparison process

The image comparison methods and algorithms usually are computation intensive and require large amount of allocated memory. As a consequence a reliable algorithm takes a lot of time to complete, much more than it is acceptable for a real time photo tagging application. Specifically, the keypoint based algorithms to the likes of the SIFT are faster if the image depicts smaller objects, as there are fewer keypoints to compare. Starting from this observation we

experimented several practically feasible methods that are simple, do not affect significantly the precision of the algorithm, but still reduce its completion time.

The first solution is to ask the user, who takes the picture at the location to select a focus area on the picture. Anyway in the majority of the cases the photo of the user depicts a larger scene than the building itself. Therefore it is not a burden for her/him to select with a rectangle the exact building/landmark to be compared. This is illustrated in *Fig.3.a*, where the image contains the building of the Parliament and the river Danube, but the building in question occupies only half of the picture. The resulting picture uploaded by the user for comparison is shown in *Fig.3.*

A second solution is to further divide the image into zones, or areas and apply the comparison process zone-wise. These two solutions should be successively applied to the same picture, as illustrated in *Fig.3.b*.



a.) ***b.)***

Figure 3: a.) Selecting the focus area within a picture b.) division of the picture into blocks.

Practically the users place the most important detail in the center of their photos. We ordered the areas of the image according to the order of processing and selected the central area of the image as the first one. In most of the photos the areas nr.8 and nr.9 are irrelevant (contain the sky), therefore those are placed at the end of the list. According to our experience the first three or four areas are enough to be processed in order to get an accurate result.

4. Performance evaluation

We tested our solution on a database of 88 images, and we compared each of them pair-wise. The images were taken in different conditions, as they would be in a real-life scenario (underexposed, overexposed, blurry). The pictures depict buildings from downtown Budapest taken from different angles and light conditions.

We have used an ASUS laptop with 1GB RAM, a 1.5GHz Intel Celeron processor and Microsoft Windows XP operating system to test our solutions. A dedicated server with faster processing units and larger memory can speed up the process. Further optimization alternative is to use a fast hard disk.

Nevertheless we considered that such a combination of database size and hardware gives a proper insight into the feasibility and performance of our proposal. In practice both the size of the database and the computational power of the server might be increased, resulting in a similar overall performance.

A. SIFT algorithm

Based on our experience the different lighting conditions, the blurred, noisy images do not significantly affect the reliability of the algorithm. On the other hand there are interfering objects in the fore- or background of the image, especially if the target object is masked (e.g., a truck or tree in front of a building).

An important issue is the angle of rotation, when the user takes his/her picture from a different angle than the reference image was registered. A similar problem is caused by the tight city street, where the optics of the mobile devices can not capture the whole front of a larger building. Even if the algorithm is scale free, these situations in practice result in lower number of captured keypoints, thus the algorithm becomes less accurate. In our case the algorithm could not recognize any building when the angle of rotation was 75° or larger.

In spite the above limitations the SIFT algorithm proves to be robust and reliable. We have found that downgraded resolution speeds up the process. We tested several solutions and we opted for a 0.5 megapixel size pictures, because the algorithm still is able to provide enough keypoints for stable operation.

We measured an 86% hit ratio and the duration of the comparison for an image was 19 seconds. We consider a successful comparison the case when the image provided by the user is compared against the images of the database and the image with largest matching keypoint pictures the same building indeed. The hit ratio expresses the ratio of successful comparisons among all the comparisons done. As a conclusion, the 86% hit ratio is an acceptable one, since in real life scenarios this can be used as the basis of a proper tagging.

The duration of a comparison is orders of magnitude larger than acceptable though. In the following we present several approaches that improved either the hit ratio or the duration of the comparison and finally we combined those that are acceptable in both terms.

B. Color based identification

The color based identification follows a different logic compared to the SIFT algorithm and its variants. This method is not a reliable one, therefore it is not used and it can not be used alone in our case. It still has a great advantage over the keypoint based solutions, since computationally it is orders of magnitude less demanding. Therefore it can be used to exclude those images that color-wise are "far" from the target image. It is also important to note that this approach is not sensible to those errors and noises that are hardly eliminated by the keypoint based algorithms. Practically the color based identification method complements the keypoint based ones.

First we applied the color based identification method in order to illustrate the above. The hit ratio has fallen to mere 56%, but the duration of a pair-wise comparison was less than 0.05 seconds. The low hit ratio confirms our expectation that this approach can not be applied alone. Due to the advantages of this approach we still did not abandon it, as shown later in this section.

C. The effect of the angle of rotation

The pictures taken from different angles decrease the accuracy of the comparison. We have found that in the case of large (i.e., above 30°) angles of rotation (the angle between the reference and the user-provided pictures) the SIFT algorithm could not take correct decisions. In order to alleviate this problem we transform the images to canonic form, and only then apply the SIFT algorithm. This method is not useful only in the case of the aforementioned large angles of rotation, since it increases the robustness of the SIFT algorithm in all cases.

We doubled the hit ratio with this method. Nevertheless this approach has several drawbacks. In the case of very large angles of rotation (above 50°) even if it increases the hit ratio, the result is still unacceptably low. Moreover, the rotation of the objects can be hardly realized in automatic manner, in most of the cases we had to adjust the automated results. Therefore further research is needed to find a solution for this transform to canonic form, which is feasible in real time environments. As a consequence in the following we do not use this transform in our tests, the presented results are obtained without this approach.

D. The effect of division of the image into areas

The main advantage of the division of the image into areas is that we can exclude those pictures right at the beginning whose low-ordered areas contain non-matching keypoints. We applied this method in combination with the SIFT algorithm and we achieved an improved hit ratio of 92%. In the same time the duration of the comparison was reduced to a mere 0.45 second for each pair-wise comparison. This value is orders of magnitude better than the 19 second achieved with SIFT alone.

E. Combination of the color based identification and the SIFT algorithm

The duration of a pair-wise comparison can be further reduced if we combine the color based identification presented earlier in this section with the SIFT algorithm.

We apply the color based identification after the approach presented in section 4.B. For each image the color based identification algorithm provides an aggregated color value. We ordered the images according to the Euclidean distances between their aggregated color values, and found that if we exclude the images from the lower half of the list we do not alter the hit ratio and at the same time we speed up the comparison process. Thus we apply the SIFT algorithm only on the half of the images compared to the method presented in section 4.B.

The effect of this approach on the duration of the comparison depends on the reliability of the color based identification method. Currently the method eliminated half of the images from further comparisons, thus the duration is approximately halved.

Theoretically this method also increases the accuracy of the comparison, since an image which significantly differs in colors from the target one is filtered out right at the beginning of the process. Nevertheless, we found that the ordered list provided by this approach in the 5% of the cases contained different buildings in the first part of the list, which lead to an increased ratio of false positives. These positive and negative effects neutralize the impact of the approach and the hit ratio is not increased. However the duration of the comparison was lowered to 0.25 seconds.

F. Summary of results

We have summarized in *Table 1* the results of our tests. Although the hit ratio of the original algorithm is already high, we could further increase it.

We can state that the processing time of a picture has been reduced by two orders of magnitude. This means that even if we compare this result against the faster SURF algorithm, the obtained gain is significant.

Table 1: Summary of the performance evaluations.

	Color based identification	*SIFT algorithm*	*SIFT algorithm + division of the image*	*SIFT algorithm + division of the image + color based filtering*
Hit ratio	56%	86%	92%	92%
Avg. processing time (per picture)	<0,05 s	19 s	0,45 s	0,25 s

If a photo tagging system is deployed in an urban environment and the location of the user is approximated as suggested in section 3.B, then the number of alternatives is around 100 pictures. Based on our tests the proposed solution keeps the response time within a single attention burst of a typical mobile user [18].

5. Conclusion

We have proposed an image identification framework solution that can support a photo tagging system. Such a photo tagging system may enable the implementation of a virtual community, social networks and related applications in urban environments.

The core of the photo tagging system is the image comparison. Our proposal builds on the widely known algorithm (SIFT) and we tried to fasten it up by finding those mechanisms that significantly reduce the per-picture processing time. Since the comparison is done at the server side, the computational resource was not a bottleneck.

We have tested our proposal on a test database. Although the transformation of the image to a canonic state significantly improves the accuracy of the process, it increases the duration of the process. A possible further research direction would be to apply this idea only in those cases when all the other mechanisms fail.

The combination of division of the image and of the color-based comparison with the original SIFT algorithm significantly improves the comparison process. At the same time it requires further research to improve the accuracy of the proposed method in the case of disturbances.

Acknowledgements

This work has been partially founded by Mobil Videó Konzorcium, Hungary. The authors would like to also thank the work of Zsolt Kosztovics and Norbert Érseki.

References

[1] Homepage of the BlueSpot project, http://bluespot.hu/.
[2] Robertson, D., Cipolla, R., "An image-based system for urban navigation" in *Proceedings of The 15th British Machine Vision Conference (BMVC'04), Kingston-upon-Thames, UK*, http://mi.eng.cam.ac.uk/reports/svr-ftp/cipolla_bmvc04.pdf, September 2004.
[3] Wang, S., Min, J., Yi, B. K., "Location Based Services for Mobiles: Technologies and Standards", *IEEE International Conference on Communication (ICC), Beijing, China*, 2008
[4] Global Positioning System, the homepage of the operator of the GPS system, http://www.gps.gov/.
[5] Varshavsky, A., et al., "Are GSM phones THE Solution for Localization?", in *Proceedings of 7th IEEE Workshop of Mobile Computing Systems and Applications (WMCSA), Semiahmoo Resort, Washington, USA*, April 2006.
[6] Mobile Codes Consortium – MC2: http://www.mobilecodes.org/.
[7] Sarfraz, S., Hellwich, O., "Head Pose Estimation in Face Recognition across Pose Scenarios", in. *Proc. of Int. conference on Computer Vision Theory and Applications, Madeira, Portugal*, pp.235-242, January 2008.
[8] Bebis, G., "Edge Detection", Department of Computer Science & Engineering, University of Nevada, USA, http://www.cse.unr.edu/~bebis/CS791E/Notes/EdgeDetection.pdf, 2003.
[9] Moeslund, T., "Canny Edge Detection", Laboratory of Computer Vision and Media Technology, Aalborg University, Denmark, http://www.cvmt.dk/education/teaching/f09/ VGIS8/ AIP/ canny_09gr820.pdf, March 2009.
[10] Kató, Z., Didactic material (in Hungarian), SZTE, Szeged, http://www.inf.u-szeged.hu/ ~kato/teaching/segmentation/03_edgedetection.pdf, July 2009.
[11] Meng, Y., Tiddeman, B., "Implementing the Scale Invariant Feature Transform (SIFT) Method", University report, University of St Andrews, St Andrews, UK, http://www.cs.st-andrews.ac.uk/~yumeng/yumeng-SIFTreport-5.18_bpt.pdf , May2006.
[12] Mikolajczyk, K., Schmid, C., "A performance evaluation of local descriptors", *IEEE Trans. on Pattern Analysis and Machine Intelligence*, Vol.10, No.27, pp.1615-1630, 2005.
[13] Bay, H., Ess, A., Tuytelaars, T., Van Gool, L., "SURF: Speeded Up Robust Features", *Computer Vision and Image Understanding (CVIU)*, Vol.110, No.3, pp.346-359, 2008.
[14] Gibara, T., "Implementation of the Canny algorithm", http://www.tomgibara.com/ computer-vision/ canny-edge-detector, 2009.
[15] Fisher, R., Perkins, S., Walker, A., Wolfart, E., "Image Processing Learning Resources", online book, http://homepages.inf.ed.ac.uk/rbf/HIPR2/hough.htm, 2004.
[16] Fisher, R., Perkins, S., Walker, A., Wolfart, E., "Affine Transformation", http://homepages. inf. ed.ac.uk/rbf/HIPR2/affine.htm, 2003.
[17] SUN Java Advanced Imaging, http://java.sun.com/javase/technologies/desktop/media/jai/.
[18] Oulasvirta, A., Tamminen, S., Roto, V., Kuorelahti, J., "Interaction in 4-second bursts: the fragmented nature of attentional resources in mobile HCI", in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, Portland, Oregon, USA*, 2005.