HAL
open science

Semi-Supervised Learning for Location Recognition from Wearable Video

Vladislavs Dovgalecs, Rémi Megret, Hazem Wannous, Yannick Berthoumieu

To cite this version:

Vladislavs Dovgalecs, Rémi Megret, Hazem Wannous, Yannick Berthoumieu. Semi-Supervised Learning for Location Recognition from Wearable Video. International Workshop on Content-Based Multimedia Indexing (CBMI), Jun 2010, Grenoble, France. 10.1109/CBMI.2010.5529903 . hal-00547964

HAL Id: hal-00547964

https://hal.science/hal-00547964v1

Submitted on 17 Dec 2010

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Semi-Supervised Learning for Location Recognition from Wearable Video

Vladislavs Dovgalecs, Rémi Megret, Hazem Wannous and Yannick Berthoumieu
Signal and Image Processing Group, IMS Laboratory, University of Bordeaux I

Abstract

This paper tackles the problem of image-based indoor location recognition. The context of the present work is activity monitoring using a wearable video camera data. Because application constraints necessitate weak supervision, a semi-supervised approach has been adopted which leverages the large amount of unlabeled images. The proposed method is based on the Bag of Features approach for image description followed by spectral dimensionality reduction in a transductive setup. Additional information from geometrical verification constraints are also considered which allowed to reach higher performance levels. The considered algorithms are compared experimentally on the data acquired in the wearable camera setup.

1 Introduction

The present study[1] is positioned in the context of video indexing of data acquired using a wearable camera. In particular, the study of lifelogs, which correspond to passive audio and video recording of one's person activities using a wearable device, shows promising perspectives in terms of retrospective memory aid [3] and task observation [2]. Small and autonomous acquisition devices such as the SenseCam device [2] now allow to record images at a rate up to one image every few seconds, during periods of several hours to several days. The amount of data to handle is therefore very large, and difficult to visualize and browse.

Our work is based on the video rate capture system presented in [1] for the monitoring and diagnosis of dementia. In this context, the visual lifelog contains the activities of a patient observed indoor at home in their usual environment. The automatic indexing of the activities is required in order to assist a practician in browsing efficiently the data to evaluate the actions and the difficulties of the patient in an ecological situation. Estimating the localization of the person within their casual environment is a prerequisite to feed higher level activities detectors with adequate contextual data.

One difficulty in our setting is the lack of supervision at the acquisition stage, which should be handled by medical aids that are not specialists of technical acquisition, and because of the little time those persons would have to devote to the modeling of the environment. We therefore expect such supervision to be done on the recorded video data itself. This would be done when taking the device into a new environment by manually annotating a fraction of the several hours long videos with the main rooms and important places. The maximum amount of information should therefore be extracted from both the labeled and unlabeled video content, which calls for a semi-supervised approach.

In this paper, we will consider a Bag of Features (BoF) method [6], which we will extend to a semi-supervised approach. Feature matching on image pairs [5] will also be examined and compared, and included into the semi-supervised approach in order to further improve the performances.

The paper is organized in several sections. Section 2 discusses related work in the domain of image recognition with an application in localization problem. Section 3 presents the proposed algorithms, which are compared experimentally and discussed in section 4.

2 Background

The goal of image based localization is to estimate the location of an unknown query image with respect to some learned database. The localization may be qualitative or quantitative. Qualitative approaches aim to estimate a 2D or 3D position whereas quantitative methods recognize a reference image or its class from the database. An example of quantitave location estimation are robotic applications as in [11, 12]. Qualitative positioning services find their application for out-

[1]This work is supported by a grant from Agence Nationale de la Recherche with reference ANR-09-BLAN-0165-02, within the IMMED project `http://immed.labri.fr/`.

doors [21, 22, 23] and indoors [24] localization using mobile devices.

Both approaches are frequently addressed by two techniques - local feature matching [5, 20, 21] with optional geometrical verification and "bag of words" model [6, 10]. Former uses geometrical constraints given local features with their positions whereas latter constructs weighted histogram of "visual words".

A large body of work attempts to extend "bag of words" model. In particular, by adding local discriminative information [25], fast location recognition from structure-from-motion point clouds [4] and finding efficiently loop closures in monocular SLAM [11]. Some works propose also to include geometrical verification [7] of query results, which was applied in [10] to refine a global "bag of words" image search to the object level.

The approach for image representation is relying on BoF visual word histograms which are known to be succesfully used in image recognition applications. While being effective, visual word histograms are typically very high dimensionality vectors. It is known that such high dimensionality spaces are very sparse and leads to well-known "curse of dimensionality" [9] or empty space phenomena problem. As we intend to learn from weak supervision, many classical machine learning methods would be prone to overfitting because of low sample number in comparison to their dimensionality. Naturally there rises the question about leveraging unlabeled images which is the subject of semi-supervised learning algorithms.

Semi-supervised learning takes into account labeled and unlabeled samples to reduce in a meaningful way the dimensionality of the problem. Dimensionality reduction (DR) implies to choose what information is preserved or how it is presented - maximum variance (PCA) [9], direction for best class separability (LDA) [9], local neighborhood preservation (LPP) [14] and others to name few. DR has been applied succesfully for face recognition [13, 17, 19] for example. Such methods are usually divided into inductive and transductive approaches.

Inductive methods learn a projective basis onto which an unseen sample could be projected (e.g. PCA, LDA). Instead, transductive learning outputs class labels for unlabeled samples directly. Methods LLE, LapEig, MDS, ISOMAP and others possess such learning framework [16].

Transductive approach, when all the data is available at the learning stage, fits perfectly our indoor localization problem and was used in our experiments.

3 Semi-supervised location recognition

We now present the proposed semi-supervised framework for location recognition. Let us represent $\mathbf{X} = (\mathbf{x}_1,\ \mathbf{x}_2,\ \ldots,\ \mathbf{x}_n)$ as a set of n image signatures, in $\Re^l$. The labeled subset of $\mathbf{X}$ represents $\mathbf{X}_l = (\mathbf{x}_1,\ \ldots,\ \mathbf{x}_l)$ with associated labels $P_l = (p_1,\ \ldots,\ p_l)$. Unlabeled samples $\mathbf{X}_u = (\mathbf{x}_{l+1},\ \ldots,\ \mathbf{x}_{l+u})$ represents the rest of the set $\mathbf{X}$ for which the labels will be estimated. In our case, the labeled data is planned to be of much smaller amount than the unlabeled data.

First, features are extracted from the images, in order to produce an affinity matrix W that contains the pairwise unsupervised visual similarities between the images of the video sequence. This affinity matrix is used in the spectral graph embedding framework to produce a reduced dimension representation of the data $\mathbf{Z}$, which is suitable for classification using standard supervised approaches.

3.1 Feature extraction

We used a BoF approach for feature extraction, based on Speeded Up Robust Features (SURF) interest point features [8] and a hierarchical k-means quantization tree [6]. This feature extraction step produces image signatures $\mathbf{X} = (\mathbf{x}_1,\ \mathbf{x_2},\ \ldots,\ \mathbf{x}_n)$ that are suitable for unsupervised similarity computation, while being efficient to compute and robust in terms of recognition. This representation can be used directly for the classification. The nearest neighbor or SVM classifiers applied on $\mathbf{X}$ will be our baseline algorithms.

3.2 Graph representation of image similarities

We define the graph $G = (V,\ E)$ where V is the set of vertices and E is the set of edges connecting neighboring vertices $(\mathbf{x}_i,\ \mathbf{x}_j)$. This graph representation encodes images as graph vertices and visual similarity information as the edges among them [14, 16]

The similarity measure s between samples $\mathbf{x}_i$ and $\mathbf{x}_j$ represent the visual similarity between the corresponding images. In our case, we used the heat kernel between the BoF image signatures ($t \in \Re$), which is generic and provided as good results as more evolved approaches in our preliminary experiments.

$$s(\mathbf{x}_i,\ \mathbf{x}_j) = e^{-\frac{\|\mathbf{x}_i - \mathbf{x}_j\|_2^2}{2t^2}} \tag{1}$$

The set of k nearest neighbors for sample $\mathbf{x}_i$ is denoted by $N_k(\mathbf{x}_i)$. Class label for sample $\mathbf{x}_i$ is denoted by $C(\mathbf{x}_i)$.

The affinity matrix is build in an unsupervised way, by keeping the similarity measures of the k-nearest-neighbours for each sample :

$$W_{BoF}(i,j) = \begin{cases} s(\mathbf{x}_i,\ \mathbf{x}_j) & \mathbf{x}_i \in N_k(\mathbf{x}_j) \text{ or } \mathbf{x}_j \in N_k(\mathbf{x}_i) \\ 0 & \text{else} \end{cases} \tag{2}$$

3.3 Spectral graph dimensionnality reduction

The general idea of the graph-embedding approach is to represent each node (original high dimensional sample $\mathbf{x}_i$) as a lower dimensionality vector $\mathbf{z}_i$ that preserves locality and relations with its neighbors encoded by graph edges. The graph Laplacian [15] is defined as $L = D - W$, where the matrix D is a diagonal matrix with values $D(i,j) = \sum_j W(i,j)$ for scaling issue elimination. In the Laplacian Eigenmap [16] approach, the optimal graph responses $\mathbf{y}^m$ are then found as the lowest eigenvectors to the following generalized eigen-problem:

$$L\mathbf{y} = \lambda D\mathbf{y} \tag{3}$$

Each graph response $\mathbf{y}^m$ represents the embedding of all samples within dimension m. Therefore, obtaining matrix $\mathbf{Y} = \left(\mathbf{y}^1, \ldots, \mathbf{y}^d\right)$ as a solution of the eigen-problem, sample $\mathbf{x}_i$ will be represented by a d dimensional vector $\mathbf{z}_i$ whose elements are comprised of ith elements of vectors $\mathbf{y}^1, \mathbf{y}^2, \ldots, \mathbf{y}^d$.

For our application all the data is available at the moment of graph construction and embedding, i.e. the learning is transductive. One of the advantages lies in the fact that the unlabeled data to be classified is used within the DR step, which allows to take into account the affinities amongst unlabeled samples in order to find the manifold on which the data actually lies.

3.4 Classification

After DR, the remaining stage of workflow remains unchanged - unlabeled sample classification. Instead of using BoF signatures $\mathbf{X}$ directly, we apply classification algorithms to their embeddings $\mathbf{Z}$. In all our experiments we used simple 1-NN classifier and Support Vector Machines (SVM) [18] classifiers.

3.5 Matching based complementary information

Following [5], we also considered pairwise bi-directional image matching based on SURF keypoints. The matches are then validated using RANSAC [7] to enforce the fundamental matrix constraint. This results in a sparse matching matrix M with the number of RANSAC validated matches. Each test image is classified with the class of the labeled image that has the largest number of validated matching features.

In order to incorporate the matching information, the matching matrix M is transformed into an affinity matrix using a non-linear ad-hoc function f. Parameters a and b define a smooth weighting of the number of matchings.

$$W_{match}(i,j) = f(M(i,j)) \tag{4}$$

$$f(m) = \begin{cases} 0 & m < a \\ \frac{m-a}{b-a} & a < m < b \\ 1 & m > b \end{cases} \tag{5}$$

The final affinity matrix used for Graph-Embedding is defined as an additive combination of BoF and Matching affinities (defined in Eq. 2 and 4):

$$W_{BM} = W_{BoF} + \alpha W_{match} \tag{6}$$

4 Experiments

The experiments will evaluate the gain of the proposed approach in the context of indoor room classification. The device presented in [1] was used to acquire the data.

4.1 Database

For evaluation purposes we recorded three independent video sequences: training, testing and a sequence for visual vocabulary construction with approximately close number of frames per class. Each video sequence depict the same six classes. It can be noted that there are strong visual similarities amongst three rooms and two corridors. In order to keep the number of training frames tractable, we obtained the testing database F1K1 by subsampling the training sequence frames (selecting each fourth image) and avoided scenes where the passage from one room to another is taking place (both for training and testing sequences). The properties of the original and subsampled databases are shown in Table 1. Some sample frames from each class are depicted in the Figure 1.

For the sake of evaluation, the labeled data will be chosen from training frame collection and evaluated using complete collection of testing sequence frames.

Sequence	Video length	Selected frames
Vocabulary Tree	14 min	21 280
Training	28 min	41 348
Testing	3 min	5 200

Database	Training samples	Testing samples
F1K1	8 629	4 845

Table 1. Experimental data description

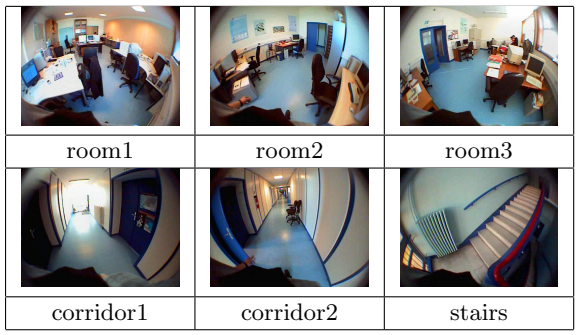

Figure 1. Content of the 6 classes

Nevertheless, both sequences will be used simultaneously at the graph-embedding step, as required by the target application.

4.2 Evaluation protocol

Each video frame is processed using the BoF framework based on SURFfeatures [5, 8]. The descriptors are quantized using a 3 levels tree with branching factor of 10 which was built using hierarchical k-means from devoted video sequence [6]. For every given frame descriptors, this produces 1111 dimensional vector (signature) All signatures are then normalized using a tf-idf scheme where the weights were computed on all the database [6].

The evaluated approaches are

1. Baseline BoF, without DR
2. BoF with linear DR such as PCA and ICA
3. BoF with graph-embedding DR in the Section 3.

Additionnally, approaches using pairwise matching were evaluated. Image similarity based on feature matching is more costly than BoF approach due to the matching algorithm. For this reason, image pairs are first screened, by considering only the first 20 nearest neigbors of each image with respect to BoF signatures.

1. Best Matching: each image is associated to the class of the image that has the largest number of matching features after RANSAC validation.
2. BoF+Matching with graph-embedding dimensionality reduction: the BoF affinities are combined with matching affinities, as explained in 6.

We evaluate the performance using accuracy measure defined as a proportion of n samples classified correctly with respect to total amount of samples N.

4.3 Baseline performance

The baseline performance was obtained using classifiers k-NN and SVM directly on the 1111 dimensions BoF signatures.

For k-NN, the best performances were obtained with $k = 1$, which is characterisic of a quite sparse sampling of the signature space.

The critical point of an SVM classifier is the choice of the kernel and selection of its parameters. The SVM classifier selected here is a soft-margin algorithm (so-called C-SVM) available online at http://www.csie.ntu.edu/ucjlin/libsvm. It has been tested with different classical kernels: linear, polynomial, radial basic function (RBF) and laplacian. We employed the parallel grid search technique combined with 5-fold cross validation to find the optimal kernel parameters for our dataset.

The Figure 3 shows the evolution of the precision when the amount of supervision is varyied by controling the number of labeled samples per class (the labeled samples are chosen randomly inside each class). 5 different sets of labeled samples were used for this experiment.

The potentially more effective SVM classifier is not able to obtain better class separation than simple $k = 1$ nearest neighbor on this dataset which is natural because of inability of SVM to avoid overfitting for such low amount of supervision given high input signature dimensionality. The signatures are therefore quite discriminative but reach a limit in the pure supervised approach. When all labeled data is used, the performances are 81.86% for 1-NN and 81.07% for SVM.

Linear unsupervised approaches PCA and ICA did not bring any significant improvement compared to direct application of k-NN or SVM and were mostly ommited from result Figure 3.

4.4 Effect of non-linear DR

The matrix W is essential for spectral graph-embedding method. We construct it providing two parameters : graph neighborhood (k) and affinity heat kernel parameter (t). Experiments showed the parameter k to be important - too low value makes the graph

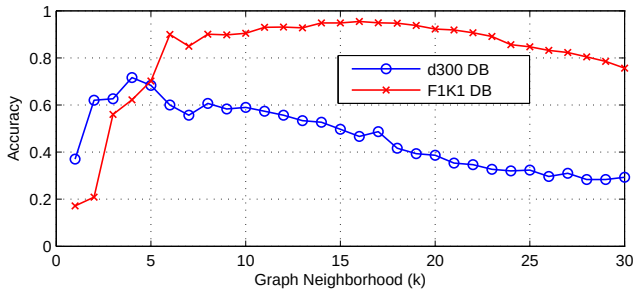

Figure 2. Influence of graph neighborhood on accuracy (for labeled 100 samples/class)

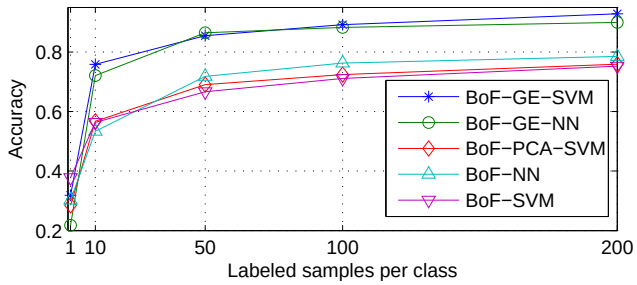

Figure 3. Influence of data preparation methods on accuracy

too sparse and even disconnected, while higher value will introduce too much irrelevant links degrading the performance. This parameter is also dependent on the database size (see Figure 2). The heat kernel parameter $t = 700$ was selected by the means of cross-validation. In figure 4, the performance of each approach is plotted with respect to the amount of supervision, expressed as a global rate of labeled samples for each class within the training samples.

Since the spectral graph approach relies on the fact that the data actually lies on a lower dimensionality manifold than the high BoF signatures, a number of dimensions has to be fixed. Figure 5 shows that the performance is optimum only for an appropriate number of reduced dimensions (between 15 and 35 dimensions). The optimal number of dimension is used for each individual measure in the Figure 4.

The graph-embedding DR applied to BoF improves the performance compared to the standard BoF approach. This improvement is more noticable for moderate amount of supervision, showing the interest of the approach. In particular, the improvement is approximately constant for a range from 100% down to 2% of labeled samples per class.

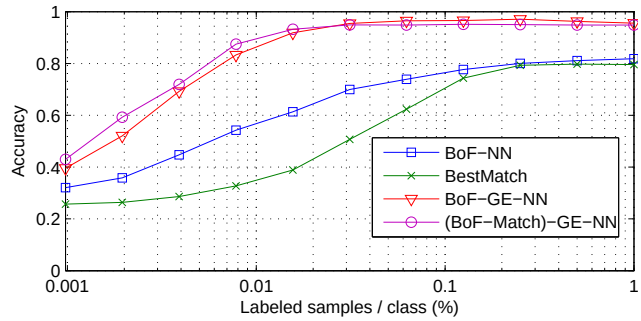

Figure 4. Comparison of BoF and Matches based approaches for various levels of supervision

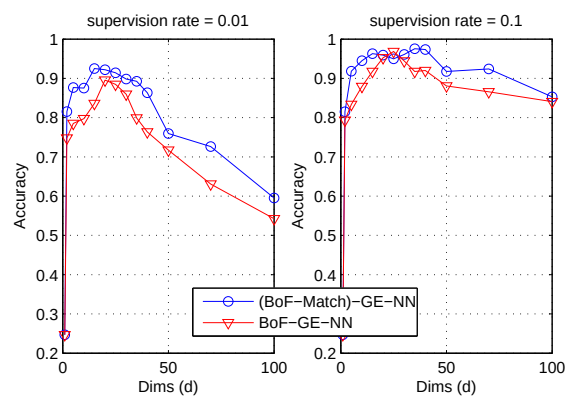

Figure 5. Intrinsic data dimensionality

The performance of keypoint matching performance is clearly decreasing faster when the amount of supervision decreases. This confirms that the matching based approach is more specific, as it can only recognize a scene that has been labeled, but can not propagate this labeling. In our case, the exact same scene may not be found in the labeled set because of the low supervision.

For the combination of BoF and matchings from eq. 6, the matches were considered to be fully reliable if more than 160 (upper threshold - b) RANSAC validated matches were found. In the other case images with a number of matches lower than 100 were considered unreliable (lower threshold - a) indicates of no match. The weighting parameter $\alpha = 0.76$ was selected manually and may be data dependent.

Results confirm the complementary nature of merging BoF and match information by demonstrating a slight increase of performance in the weak supervision situations.

5 Conclusion

In this paper we address indoor localization problem from video sequence recorded by a camera wearer. Our method is capable of learning from low amounts of labeled data - down to a few percents from the total amount. Experiments showed our approach to be efficient even in presence of visual ambiguities between classes.

Indeed, Graph-Embedding DR with inclusion of matching information results in an increase of the performance over plain BoF approach, especially for weaker levels of supervision.

Additionally, RANSAC validated matches alone showed to be insufficient for the present task because of the high subsampling rate of the labeled data, but it provides reliable information that was shown to further improve the results of the BoF approach within the semi-supervised framework. In the future, complementary data such as contextual information or the output of inertial sensors, may be integrated in a similar way in order to increase the discriminative power. Larger scale acquisition campaign on volunteers in real conditions are currently planned as a part of the IMMED project in order to evaluate the algorithms on a larger variety of conditions.

References

[1] R. Mégret, D. Szolgay, J. Benois-Pineau, P. Joly, J. Pinquier, J.-F. Dartigues and C. Helmer. "Wearable Video Monitoring of People With Age Dementia: Video Indexing at the Service of Healthcare ", CBMI 2008, London.

[2] Byrne D, Doherty A.R, Smeaton A.F, Kumpulainen S and Jarvelin K. The SenseCam as a Tool for Task Observation. . HCI 2008 – 22nd BCS HCI Group Conference, Liverpool, U.K., 1-5 September 2008.

[3] E. Berry, N. Kapur, L. Williams, S. Hodges, P. Watson, G. Smyth, J. Srinivasan, R. Smith, B. Wilson, and K. Wood, "The use of a wearable camera, SenseCam, as a pictorial diary to improve autobiographical memory in a patient with limbic encephalitis," Neuropsychological Rehabilitation, vol. 17, numbers 4/5, pp. 582-681, August 2007.

[4] A. Irschara et al., "From Structure-from-Motion Point Clouds to Fast Location Recognition", CVPR 2009.

[5] C. O. Connaire, M. Blighe and N. E. O'Connor, "SenseCam Image Localisation using Hierarchical SURF Trees", 15th International Multimedia Conference, 2009.

[6] D. Nister and H. Stewenius, "Scalable Recognition with a Vocabulary Tree", CVPR 2006.

[7] R. Raguram et al., "A Comparative Analysis of RANSAC Techniques to Adaptive Real-Time Random Sample Consensus", In Proc. ECCV 2008.

[8] H. Bay et al., "SURF : Speeded Up Robust Features", CVIU, Vol. 110, No. 3, pp. 346-359, 2008

[9] R. O. Duda, P. E. Hart and D. G. Stork, "Pattern Classification, 2nd Edition", Wiley, November 2000.

[10] J. Sivic and A. Zissermann, "Video Google : A Text Retrieval Approach to Object Matching in Videos", Proc. of the International Conference on Computer Vision, 2003.

[11] E. Eade and T. Drummond, "Unified Loop Closing and Recovery for Real-Time Monocular SLAM", BVMC, September 2008.

[12] G. Schindler et al., "City-Scale Location Recognition", IEEE Conference in Computer Vision and Pattern Recognition, 2007.

[13] M. Turk and A. P. Pentland, "Face Recognition using Eigenfaces", IEEE CVPR, Maui, Hawaii, 1991.

[14] X. He and P. Niyogi, "Locality Preserving Projections", Cambridge, MA : MIT Press, 2004.

[15] Fan R. K. Chung, "Spectral Graph Theory", Regional Conference Series in Mathematics, number 92, 1997.

[16] M. Belkin and P. Niyogi, "Laplacian Eigenmaps and Spectral Techniques for Embedding and Clustering", NIPS, Vancouver, British Columbia, Canada, 2002.

[17] X. He et al., "Face Recognition using Laplacianfaces", IEEE Trans. on Pattern Analysis and Machine Intelligence, Vol. 27., No. 3, pp. 328-340, 2005.

[18] I. Steinwart and A. Christmann, "Support Vector Machines", Springer-Verlag, New York, 2008.

[19] P.N. Belhumeur, J.P. Hespanha, D.J. Kriegman, Eigenfaces vs. Fisherfaces: recognition using class specific linear projection, IEEE Trans. Pattern Anal. Mach. Intell. 19 (7) (1997) 711–720.

[20] N. Yazawa et al., "Image Based View Localization System Retrieving from a Panorama Database by SURF", IAPR Conference MVA2009, pp. 118-121, May 18-22, 2009.

[21] C. Valgren and A. Lilienthal, "SIFT, SURF and seasons : Long-term outdoor localization using local features", In. Proc. of 3rd European Conference on Mobile Robots, Freiburg, 2007.

[22] W. Zhang and J. Kosecka, "Image Based Localization in Urban Environments", Int. Symposium on 3D Data Processing, Visualization and Transmission, 3DPVT 2006, North Carolina, Chapel Hill.

[23] T. Yeh et al., "Searching the web with mobile images for location recognition", In : CVPR, 2004.

[24] N. Ravi et al., "Indoor Localization Using Camera Phones", In WMCSA '06 : Proc. of 7th IEEE Workshop on Mobile Computing Systems, 2006.

[25] A. Quattoni and A. Torralba, "Recognizing Indoor Scenes", IEEE Conference on Computer Vision, 2009.