EXEMPLAR BASED METRIC LEARNING FOR ROBUST VISUAL LOCALIZATION

C. Le Barz *N. Thome, M. Cord* *S. Herbin, M. Sanfourche*

Thales Theresis, 91767 Palaiseau, France
Sorbonne University UPMC Univ. Paris 06, 75005 Paris, France
ONERA - The French Aerospace Lab, 91123 Palaiseau, France

ABSTRACT

This paper presents an exemplar based metric learning framework dedicated to robust visual localization in complex scenes, *e.g.* street images. The proposed framework learns off-line a specific (local) metric for each image of the database, so that the distance between a database image and a query image representing the same scene is smaller than the distance between the current image and other images of the database. To achieve this goal, we generate geometric and photometric transformations as proxies for query images. From the generated constraints, the learning problem is cast as a convex optimization problem over the cone of positive semi-definite matrices, which is efficiently solved using a projected gradient descent scheme. Successful experiments, conducted using a freely available geo-referenced image database, reveal that the proposed method significantly improves results over the metric in the input space, while being as efficient at test time. In addition, we show that the model learns discriminating features for the localization task, and is able to gain invariance to meaningful transformations.

Index Terms— content-based image retrieval, supervised metric learning, visual localization, place recognition

1. CONTEXT

The problem tackled in this paper is visual localization at a street level [1] [2] [3]. Visual localization methods differ from each other by the type of features extracted online (2D only [2] and/or 3D [4]), the matching method (dedicated or not to large scale [5] [6]), and the a priori information used (geo-referenced database image [2], 3D model [3], 2D roadmap [7], mobile cell phone information [6],...). Our typical scenario is the precise localization of a vehicle whose approximate localization is known. Our problem is cast to an image retrieval (IR) problem using only 2D features extracted from acquired images and 2D geo-referenced database image (Fig. 1). Exploiting only image content is challenging because, even if query and database images depict the same scene, camera-view points, illumination and colorimetry are different, the scene itself may have changed over time and been occluded. Preliminary experiments made clear that standard image retrieval approaches may not always be selective enough for street areas because the same features tend to be shared by several neighbour images. Standard matching methods can be classified into voting-based strategies [8] and methods relying on the Bag of Words (BOW) model [9]. Voting-based methods [8] search for each query image descriptor the N nearest descriptors belonging to database descriptors. Each of these N nearest descriptors votes for a database image. The images having the highest vote number are likely to be similar images. Ultimately, geometrical verification is often used to further improve performances. These methods are effective, but are very time consuming and do not scale well to large databases. BOW-based methods [9] quantify local descriptors of images with a codebook of visual words to generate a visual words histogram. The codebook is previously learned by clustering feature space of a learning image database in K visual words. Several extensions have been proposed, in terms of coding (Vector of Locally Aggregated Descriptors (VLAD) [10] or Fischer Vector (FV) [11]), or pooling [12], or integration of spatial information [13]. BOW-based methods are fast, and have further been optimized to find images in huge database images.

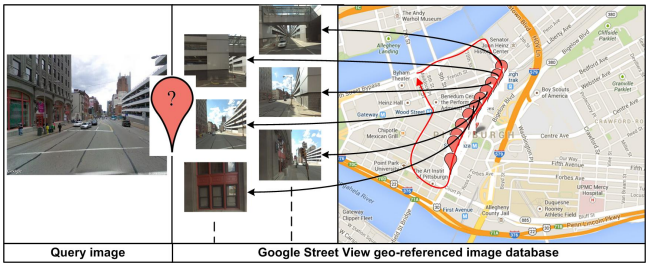

Fig. 1. Our system aims at answering the following question: knowing a rough position of the vehicle in a street and the scene being observed by the vehicle's camera, can we determine where is it exactly along the street?

The choice of the visual similarity is a key ingredient for effective image retrieval systems. A common choice is the Euclidean or χ^2 distance to find the images that are most sim-

ilar to the query image. Another appealing solution is metric learning, which proves to be useful for many image processing tasks, such as image classification, retrieval or face verification. Metric learning algorithms learn a transformation of data which is optimized for a prediction task, such as nearest-neighbour classification. The metric learning literature is very abundant, and an exhaustive review of existing methods is outside the scope of the paper. The reader can refer to [14]. Basically, methods can be classified according to the type of training data, *e.g.* pairs [15], triplets [16] or quadruplets [17], and the distance parametrization.

In this paper, we propose an exemplar-based metric learning framework, Exabal, dedicated to robust visual localization in complex scenes such as street images. Our solution is dedicated to street scale localization instead of city scale localization as proposed in [18]. The rationale of Exabal is to learn a local pseudo-distance matrix for each image of the database, so that the distance between a database image and the given query acquired from the same scene is smaller than the distance between this query and other (close) images in the database. To achieve this goal, our method encompasses the following contributions:

- During training, we generate sensible geometric and photometric transformations to model images similar to unknown query images. This offers the possibility to learn features able to discriminate a given dataset image from its neighbours, and at the same time to learn invariance to common transformations occurring at test time.
- The learning of a local metric for each dataset image is cast as a convex optimization problem, which is efficiently solved with a projected gradient descent scheme. Once this off-line procedure is carried out, computing the similarity at test time is very efficient, *e.g.* much faster than methods based on kNN votes like [1]. It can also benefits from existing fast indexing structures (*e.g.* inverted files, search trees).
- Successful experiments reveal the ability of the method to effectively retrieve images in complex street scene. We show that the model learns sensible invariances and discriminating features for localization.

2. EXABAL METHOD FOR METRIC LEARNING

We consider here the widely used Mahalanobis distance metric $D_{\mathbf{M}}$ that is parametrized by the positive semi-definite matrix (PSD matrix) $\mathbf{M} \in \mathbb{S}_+^d$ such that the distance between vectorial representations $(\mathbf{x}_j, \mathbf{x}_k) \in \mathbb{R}^d \times \mathbb{R}^d$ of the images (I_j,I_k) is written as follows:

$$\mathcal{D}^2_{\mathbf{M}}(\mathbf{x}_j, \mathbf{x}_k) = (\mathbf{x}_j - \mathbf{x}_k)^\top \mathbf{M} (\mathbf{x}_j - \mathbf{x}_k) \qquad (1)$$

Learning a pseudo-distance matrix using Eq. (1) is equivalent to learning a linear transformation of data, since any PSD matrix M can be decomposed as: $\mathbf{M} = \mathbf{L}^\top \mathbf{L}$ where $\mathbf{L} \in \mathbb{R}^{e \times d}$.

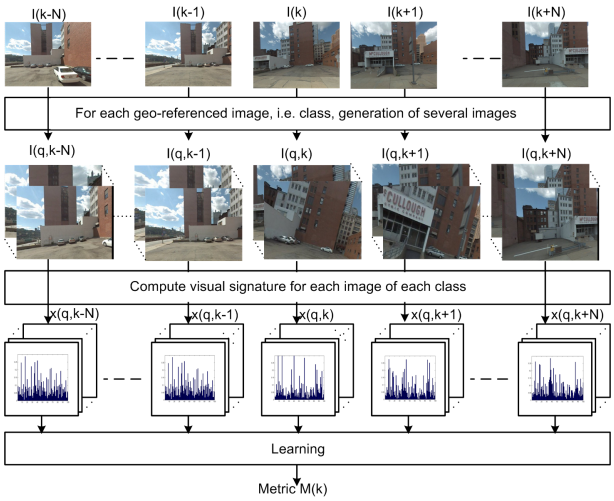

Fig. 2. Creation of simulated similar and dissimilar examples in order to learn for each geo-referenced image database I_k a metric $\mathbf{M}_k$.

The method can be easily extended to non-linear metric learning using kernel methods [14].

The overall pipeline of the proposed metric learning scheme is illustrated in Figure 2. The rationale of the method is to learn a local distance matrix $\mathbf{M}_k$, as defined in Eq. (1), for each image of the database I_k, leading to an exemplar-based metric learning scheme.

2.1. Exemplar-Based Constraints

Basically, we require that the distance between each image signature $\mathbf{x}_k$ and other neighbour image signatures of the database $\mathbf{x}_{k'}$, $k' \neq k$, is larger than the distance between $\mathbf{x}_k$ and a query signature $\mathbf{x}_q$ representing the same scene as I_k. During training, the main challenge is to find images which are representative of the unknown query images that we have to localize. To achieve this goal, we propose to generate geometric and photometric transformations from I_k as proxies for potential test query images.

Formally, let us denote as $\mathcal{T}$ the set of considered transformations. For $(T_s^{(i)}, T_d^{(i)}) \in \mathcal{T} \times \mathcal{T}$, we obtain the transformed images $T_s^{(i)}(I_k)$ and $T_d^{(i)}(I_{k'})$ for I_k and $I_{k'}$, respectively. We note $T_s(\mathbf{x}_k)$ and $T_d(\mathbf{x}_{k'})$ the vectorial signature of $T_s^{(i)}(I_k)$ and $T_d^{(i)}(I_{k'})$, respectively. During training, we enforce the following constraint:

$$\mathcal{D}_{\mathbf{M}_k}(\mathbf{x}_k, T_d(\mathbf{x}_{k'})) \geq \mathcal{D}_{\mathbf{M}_k}(\mathbf{x}_k, T_s(\mathbf{x}_k)) + 1 \qquad (2)$$

The constraint in Eq. (2) promotes matrices $\mathbf{M}_k$ that discriminate I_k images from $I_{k'}$ images, at the same time as taking into account potential transformations. An appealing property of our constraint generation approach is the ability to produce a large number of constraints by sampling different $T_s^{(i)}$ and $T_d^{(i)}$ transformations, making the optimization of $\mathbf{M}_k$ (with a potentially large number of parameters) robust to over fitting.

In this paper, we focus on rotations and cropping operations, but $\mathcal{T}$ could easily been enriched if required. Each image I_k is described by a feature $\mathbf{x}_k$ corresponding to a BOW vector [9] encoding spatial information [13]. We validate in the experiments (section 3) that the method learn a distance that selects discriminative and spatially localized features, making the similarity measure much more powerful than the distance in the input space.

2.2. Optimization

The objective of our learning scheme is to minimize the number of misclassified constraints in Eq. (2). Since the direct resolution of this problem is NP-hard, we introduce a standard hinge loss function ℓ_d for penalizing the violation of each constraint in Eq. (2): $\ell_d(\mathbf{x}_k, T_s(\mathbf{x}_k), T_d(\mathbf{x}_{k'})) = max\left[0, 1 - (\mathcal{D}_{\mathbf{M}_k}(\mathbf{x}_k, T_d(\mathbf{x}_{k'})) - D_{\mathbf{M}_k}(\mathbf{x}_k, T_s(\mathbf{x}_k)))\right]$.

In addition, we incorporate into our objective function the following convex loss ℓ_s for each pair $(I_k, T_s^{(i)}(I_k))$: $\ell_s(\mathbf{x}_k, T_s(\mathbf{x}_k)) = \mathcal{D}_{\mathbf{M}_k}(\mathbf{x}_k, T_s(\mathbf{x}_k))$. ℓ_s aims at minimizing the distance between each image and its transformed version, *i.e.* between similar images. As in [16], it can be interpreted as a regularization prior. Note that other regularization schemes could also be used, *e.g.* based on the Frobenius or nuclear norm [19].

Our final objective $\mathcal{P}(\mathbf{M}_k)$ function combines the loss ℓ_s and ℓ_d over the whole set of constraints, with a weighting parameter μ:

$$\mathcal{P}(\mathbf{M}_k) = (1-\mu) \sum_{T_s \in \mathcal{T}} \ell_s(\mathbf{x}_k, T_s(\mathbf{x}_k)) + \mu \sum_{\substack{k' \neq k, \\ (T_d, T_s) \in \mathcal{T} \times \mathcal{T}}} \ell_d(\mathbf{x}_k, T_s(\mathbf{x}_k), T_d(\mathbf{x}_{k'})) \quad (3)$$

The objective function in Eq. (3) is convex with respect to $\mathbf{M}_k$. To solve it, we use a stochastic projected gradient descent scheme. After each gradient computation[1], the matrix $\mathbf{M}_k$ is updated and projected onto the PSD cone if necessary. The algorithm is guaranteed to converge to the global minimum, up to a well-chosen gradient step. In practise, the optimization is fast with reasonable number of constraints and quickly converges.

2.3. Localization Model

Once a metric $\mathbf{M}_k$ is learned for each image I_k of the database, the goal of our system at test time is to localize a given query image I_q, with vectorial representation $\mathbf{x}_q$. To achieve this goal, we assume here that we want to match I_q against N given images of the database $\{I_1, ..., I_N\}$ (for example selected using GPS information). Our system thus output the image I_{k^*}, which is the closest to I_q, as follows :

$$k^* = \underset{k \in \{1;N\}}{\arg\min}\ D_{\mathbf{M}_k}(\mathbf{x}_k, \mathbf{x}_q) \quad (4)$$

[1]This computation is easy since each term is (piece-wise) linear in $\mathbf{M}_k$.

where $D_{\mathbf{M}_k}(\mathbf{x}_k, \mathbf{x}_q)$ is the distance between representations of I_q and I_k images computed with the matrix $\mathbf{M}_k$, obtained after training as described in section 2.2. When computing the minimization in Eq (4), there is no obvious guarantee that the different distances $D_{\mathbf{M}_k}(\mathbf{x}_k, \mathbf{x}_q)$ for different $\mathbf{M}_k$ are comparable, since each optimization has been performed independently. To alleviate this problem, we normalize each $\mathbf{M}_k$, as a post-processing learning step, so that the Frobenius norm of $\mathbf{M}_k$ is equal to 1. We could use more advanced normalization schemes as proposed in [20], but we found it sufficient in our experiments.

3. EXPERIMENTAL RESULTS

3.1. Experimental setup

We built an image corpus from Google Pittsburgh dataset [21] for image database, and from Google Streetview images for query images [22]. These image dataset have been acquired at different time, resulting in strong visual changes for the same scenes. Camera fields of view are also different. From the original corpus, we keep one image every 5m resulting in a corpus of 2215 images. Query images are downloaded from Google Streetview website (resolution of 640x480, field of view of 100^o, camera tilt of 5^o.) We requested one image every 15m resulting in 846 query images.

BOW are computed from SIFT descriptors densely extracted: four scales are used 1, 1.5, 2, 2.5 and the step between each descriptor is 4. BOW parameters are the followings: Hard assignment, Sum pooling, L2 Normalization, no tf-idf weighting. Spatial Pyramidal Matching configuration for BOW is 1x1, 2x2. The size of the codebook has been chosen to be 100. The trade-off parameter μ between the two terms ℓ_s and ℓ_d of the objective function, was set to 0.5. Image retrieval is performed among N=11 database images, which is equivalent to a localization uncertainty of 50m.

We compared our solution (Exabal) with state of the art IR solutions: a BOW based method with spatial pyramidal matching (BOW), and a kNN votes method like [1] with and without a geometric consistency verification, respectively noted (kNN+RANSAC) and (kNN).

3.2. Results

Achieved performances using the previously described setup, are given in Tab. 1. These results show the interest of exemplar based local metric learning for visual localization.

Compared to (BOW) method, our solution achieves a substantial gain of 12% (from 48% to 60%). Fig.3 presents some examples where Exabal makes it possible to find the database image (b) that depicts the same scene as the query image (a), whereas (BOW) solution can't (c). Achieved performances validate our objective function defined by Eq. 3, as well as the way we generate the constraints, *i.e.* by applying various

Method	Loc. error	Accuracy	Time
(kNN)	9.5m	46%	16.1s
(kNN+RANSAC)	6.0m	58%	21.5s
(BOW)	9.6m	48%	2.1s
(Exabal)	5.8m	60%	3.0s

Table 1. Mean localisation error, accuracy and time to process one query for state of the art IR algorithms based on Euclidean distance for visual similarity and for our solution (Exabal) based on local learned Mahalanobis distance.

transformations to database images in order to build representative images that likely look like to potential query images. The processing time is slightly higher as the Mahalanobis distance is more complex to compute than an Euclidean distance.

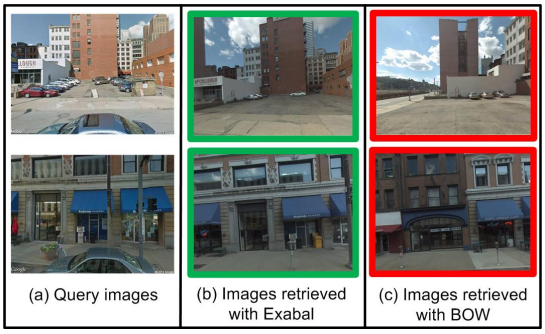

Fig. 3. (a) Query images - (b) Images retrieved with (Exabal) solution - (c) Images retrieved with (BOW) solution.

Compared to effective but time consuming methods, the proposed solution achieved better performances than (kNN) (from 46% to 58%) and is even slightly better than (kNN+RANSAC) (from 58% to 60%). At the same time, our method is faster: the processing time to process one query (measured with Matlab) is roughly reduced by a 10 factor.

Thereafter, we analyse the reasons of performance improvements. We show that they are due to the selection of discriminative features, the selection of discriminative image area(s) and, the learning of invariance to photometric and geometric transformations.

- Selection of discriminative features and discriminative image area(s)

The first advantage of our method is that it selects relevant, *i.e.* discriminative visual features. To visualize the most discriminative word for a given database image, we compute the eigenvector $\mathbf{v}_1$ of the largest eigenvalue λ_1 of $\mathbf{M}_k$ that represents the importance of each visual word[2]. Fig. 4 shows the visual word having the highest value in vector $\mathbf{v}_1$. This visual word "window corner" makes it possible to retrieve the good image (b), although many features (the bricks of the wall) were common between the query image (a) and a neighbour image (c). What's more, as the BOW hold spatial information, we also check that Exabal is able to learn where the most discriminative features are located in the image.

[2] $\mathbf{M}' = \lambda_1 \mathbf{v}_1 \mathbf{v}_1^T$ is the nearest rank-1 matrix of $\mathbf{M}$ in the ℓ_2 norm. Thus $D_M^2(x_j, x_k) \approx \lambda_1 (\mathbf{v}_1^T (x_j - x_k))^2$ and therefore $\mathbf{v}_1$ weights the importance of visual words.

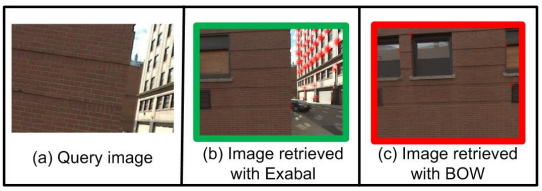

Fig. 4. The visual word "window corner", that has been learned to be discriminating, improves image retrieval task.

- Learning invariance to photometric and geometric transformations

An other advantage of our method is that we learn invariance to photometric and geometric transformations. To demonstrate it, we selected randomly 1000 database images on which various transformations have been applied (*i.e.* various rotations from $-18°$ to $+18°$ around the 3 axes and various crops from 6 to 35 pixels). Thus we generate 10000 query test images $I_q = T^{(i)}(I_k)$, that have not been used during training. The mean classification rates for these simulated query images are reported in Tab. 2 for (BOW) and (Exabal) solutions. A classification rate of more than 99% when learned Mahalanobis distance is used confirms our claim concerning invariance to photometric and geometric transformations.

	Mean classification rates
(BOW)	94.8%
(Exabal)	99.1%

Table 2. Mean classification rates computed for simulated test images.

4. CONCLUSION

We proposed a new image retrieval framework dedicated to applications for which geographical position (GPS) of database images are available as well as an *a priori* approximate localization of the query image. The proposed solution is simple and fast: for each query image, only one BOW has to be computed and N Mahalanobis distances. The proposed framework can also benefit from existing fast indexing methods. We compared our framework with state of the art image retrieval algorithms evaluated on a corpus of 846 queries and 2215 database images. Our solution improves accuracy from 48% for a traditional BOW solution to 60%, while maintaining the same processing time. We plan to incorporate this new framework improving local visual similarity measurements in the system exploiting spatio-temporal constraints induced by vehicle moving as proposed in [23]. Finally, even if we use BOW as visual signature, other visual signatures can be easily used, as the recently deep features [24] which have been demonstrated to be efficient.

5. REFERENCES

[1] A. Zamir and M. Shah, "Accurate image localization based on google maps street view," in *Proceedings of the European Conference on Computer Vision*, Sept. 2010, pp. 255–268.

[2] G. Vaca-Castano, A.R. Zamir, and M. Shah, "City scale geo-spatial trajectory estimation of a moving camera," in *Proceedings of the Computer Vision and Pattern Recognition conference*, 2012, pp. 1186–1193.

[3] H. Liu, T. Mei, J. Luo, H. Li, and S. Li, "Finding perfect rendezvous on the go: accurate mobile visual localization and its applications to routing," in *Proceedings of the ACM Multimedia Conference*, Oct. 2012, pp. 9–18.

[4] H. Badino, D. Huber, and T. Kanade, "Real-time topometric localization," in *Proceedings of the International Conference on Robotics and Automation*, May 2012, pp. 1635–1642.

[5] G. Schindler, M. Brown, and R. Szeliski, "City scale location recognition," in *Proceedings of the Computer Vision and Pattern Recognition conference*, June 2007, pp. 1–7.

[6] J. Zhang, A. Hallquist, E. Liang, and A. Zakhor, "Location-based image retrieval for urban environments," in *Proceedings of the International Conference on Image Processing*. IEEE, 2011, pp. 3677–3680.

[7] M.A. Brubaker, A. Geiger, and R. Urtasun, "Lost! leveraging the crowd for probabilistic visual self-localization," in *Proceedings of the Conference on Computer Vision and Pattern Recognition*, June 2013, pp. 3057–3064.

[8] D.G. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 60, pp. 91–110, nov 2004.

[9] J. Sivic and A. Zisserman, "Video Google: A text retrieval approach to object matching in videos," in *Proceedings of the International Conference on Computer Vision*, Oct. 2003, vol. 2, pp. 1470–1477.

[10] H. Jégou, M. Douze, P. Pérez, and C. Schmid, "Aggregating local descriptors into a compact image representation," in *Proceedings of the Conference on Computer Vision and Pattern Recognition*, June 2010, pp. 3304–3311.

[11] F.Perronnin and C. Dance, "Fisher kernels on visual vocabularies for image categorization," in *Proceedings of the Conference on Computer Vision and Pattern Recognition*, June 2007, pp. 1–8.

[12] Sandra Avila, Nicolas Thome, Matthieu Cord, Eduardo Valle, and Arnaldo de Albuquerque Araújo, "BOSSA: extended BoW formalism for image classification," in *18th IEEE International Conference on Image Processing (ICIP 2011)*, Sept. 2011, pp. 2966–2969.

[13] S. Lazebnik, C. Schmid, and J. Ponce, "Beyond bag of features: spatial pyramid matching for recognizing natural scene categories," in *Proceedings of the Computer Conference on Computer Vision and Pattern Recognition*, June 2006, pp. 2169–2178.

[14] B. Kulis, "Metric learning: A survey," *Foundations and Trends in Machine Learning*, vol. 5, pp. 287–364, 2013.

[15] E.P. Xing, A.Y. Ng, M.I. Jordan, and S. Russell, "Distance metric learning, with application to clustering with side-information," in *NIPS*, 2002.

[16] K. Weinberger and L. Saul, "Distance metric learning for large margin nearest neighbor classification," *In the Journal of Machine Learning Research*, vol. 10, pp. 1453–1484, June 2009.

[17] Marc Teva Law, Nicolas Thome, and Matthieu Cord, "Quadruplet-wise image similarity learning," in *ICCV*, 2013.

[18] P. Gronat, G. Obozinski, J. Sivic, and T. Pajdla, "Learning and calibrating per-location classifiers for visual place recognition," in *Proceedings of the Conference on Computer Vision and Pattern Recognition*, June 2013.

[19] B. McFee and G. R. G. Lanckriet, "Metric learning to rank," in *ICML*, 2010.

[20] M. Gebel and C. Weihs, *Calibrating classifier scores into probabilities*, Advances in Data Analysis. Springer Science and Business Media, 2007.

[21] Google company, "Pittsburgh dataset provided by google for research purposes," `http://www.icmla-conference.org/icmla11/challenge.html`.

[22] Google company, "Google street view API," `http://developers.google.com/maps/documentation/streetview`.

[23] C. Le Barz, N. Thome, M. Cord, S. Herbin, and M. Sanfourche, "Global robot ego-localization combining image retrieval and hmm-based filtering," in *6th workshop on Planning Perception and Navigation for Autonomous Navigation*, Sept. 2014.

[24] S. Razavian, H. Azizpour, J. Sullivan, and S. Carlsson, "Cnn features off-the-shelf : An astounding baseline for recognition," in *Proceedings of Computer Vision and Pattern Recognition conference*, 2014.