SCENE AND VIEWPOINT BASED VISUAL SUMMARIZATION FOR LANDMARKS

Weiqing Min[1], *Bing-Kun Bao*[1,2], *Changsheng Xu*[1,2]

[1]National Lab of Pattern Recognition, Institute of Automation, CAS, Beijing
[2]China-Singapore Institute of Digital Media, Singapore

ABSTRACT

Visual summarization of landmarks is an important task for applications, such as landmark organization, search and browsing. In this work, we make the first attempt towards landmark summarization by simultaneously considering both the scenes (e.g., sunny view and night view) and viewpoints (e.g., front-side and close-distant viewpoint). In the proposed framework of landmark summarization, we first group images into different clusters by viewpoints, then the distinctive scenes for each viewpoint cluster are discovered by the proposed scene-viewpoint based theme modeling. Compared with the existing topic models, our model is capable of mining scene-viewpoint themes directly from all viewpoint clusters and meanwhile differentiating among these themes by viewpoints. The landmark summary is generated by the discovered scene-viewpoint themes, where each theme is represented by the selected images with one certain scene and viewpoint. The experimental results validate the proposed method and demonstrate its advantage in improving user experience.

Index Terms— Landmark Summary, Scene Theme, Viewpoint, Theme Modeling

1. INTRODUCTION

With the popularization of online image-sharing services like Flickr, there are a huge number of landmark images available. As shown in Fig.1(a), these images are different in both scenes (e.g. cloudy view and sunny view) and viewpoints (e.g. front-side viewpoint, close-distant viewpoint, etc). A series of landmark images with diverse scenes and viewpoints can comprehensively summarize a landmark with low redundancy. Such summary can help to better organize and browse image collections of a particular landmark. Furthermore, the summarized results have many potential applications such as landmark recognition [1] [2] , landmark retrieval [3] [4] [5] [6] and automatic 3D reconstruction [7].

Much of existing work has been devoted to visual summarization on landmarks only based on the scenic view [8] [9] [10] [11] [12] or viewpoint [13]. For scenic view based landmark summary, Simon et al. [8] focused on visual clustering to find a set of canonical views for constructing the visual scene. Kennedy et al. [9] moved one step further and used both context and content information, such as tags and visual features to summarize representative views of landmarks. In contrast, Zhao et al. [10] defined the scenic theme as a distinct landmark scenic view and proposed to summarize landmarks by the scenic theme. Different from [8] [9] [10], Xue et al. [13] modeled an image's viewpoint approximately by detecting the same part of an object in different positions, scales and orientations. However, all the existing work is probably unsatisfactory to users for not providing comprehensive views of landmarks, as they would prefer the landmark images with diverse viewpoints and scenes.

Therefore, we claim that the visual summary of one landmark should involve the following two aspects: (1) **Scene**. One landmark presents different scenic views under different time of day or in different weather conditions, such as the sunset view for Big Ben or the night view for Eiffel Tower; (2) **Viewpoint**. Images taken from different viewpoints probably present larger differences in content. To better summarize the landmarks, the purpose of our work is to discover underlying scene-viewpoint themes from the landmark images.

As the first attempt, the main challenge to our work is the inconsistence in the representations of scenes and viewpoints, that is, the scenes are represented by the appearance information of images while the viewpoint is represented by the geometry information. This motivates our ideas to split our framework into two stages. The first stage is to cluster images with similar viewpoints together by viewpoint clustering, and the second one is to discover distinctive scenes from each viewpoint cluster via scene-viewpoint based theme modeling. When an image is represented by Bag-of-Words, the landmark images with the same scene should share the same visual words, which are distinctive from others. Therefore, we can utilize the topic model to discover these distinctive visual words for scenes. Compared with the existing topic models, our model is capable of simultaneously discovering the scene-viewpoint themes directly from all viewpoint clusters and differentiating among these themes by viewpoints. With the discovered scene-viewpoint themes, we can easily create a summary for a given landmark. Fig.1(b) shows the summarized results for Big Ben. We can see that this landmark is summarized by scene-viewpoint themes, where each theme is represented by one image with one scene (corresponding to

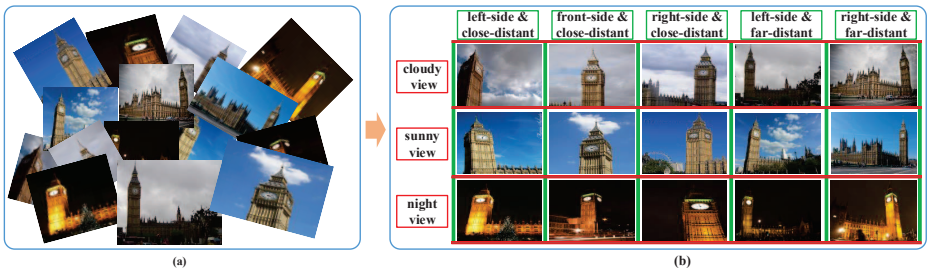

Fig. 1: Visual summary for Big Ben

one row) and viewpoint (corresponding to one column).

2. VIEWPOINT CLUSTERING

Raguram et al. [7] considered that the viewpoint of one landmark is determined by the spatial distribution of the landmark in an image and proved that the global descriptor GIST is effective for grouping images of the same landmark into clusters with similar viewpoints, Inspired by this work, we also select the GIST features and construct the feature similarity matrix , where the visual similarity $W_{p,q}$ between two images can be computed based on the Gaussian kernel function with a radius parameter σ, i.e.,

$$W_{p,q} = exp(-\frac{||\mathbf{v}_p - \mathbf{v}_q||^2}{\sigma^2}) \quad (1)$$

where $\mathbf{v}_p$ and $\mathbf{v}_q$ denote the feature vector of image p and q, respectively (in our experiment, $\sigma = 0.15$).

We then perform the spectral clustering [14] to cluster images with similar viewpoints using the constructed similarity matrix. After spectral clustering on images, we obtain a set of visual clusters $\{G_{l,1}, ..., G_{l,k}, ...G_{l,|K|}\}$ for each landmark l. In order to identify representative viewpoint clusters and filter out inconsistent clusters, we next rank viewpoint clusters. Particularly, we compute average intra-cluster similarity S_{intra} for each cluster k, which is defined as the average visual similarity between images within the cluster, i.e.,

$$S_{intra} = \frac{1}{|G_{l,k}|} \sum_{p \in G_{l,k}} \frac{1}{|G_{l,k}| - 1} \sum_{q \in G_{l,k} \setminus p} W_{p,q} \quad (2)$$

A high intra-cluster similarity S_{intra} indicates that the cluster is more tightly formed and representative.

Finally, we select clusters with high intra-cluster similarity for next theme modeling. Fig. 2 shows top 4 ranked clusters for Big Ben and Arc de Triomphe, respectively. Each column represents one viewpoint cluster and each image p from one viewpoint cluster k is ranked by the degree D_p

$$D_p = \sum_{q \in G_{l,k}} W_{p,q} \quad (3)$$

3. SCENE-VIEWPOINT BASED THEME MODELING

Obviously, each viewpoint cluster includes different scenes. As shown in Fig.2, the images of Arc de triomphe with the front view (the second row) consist of sunset, night, sunny

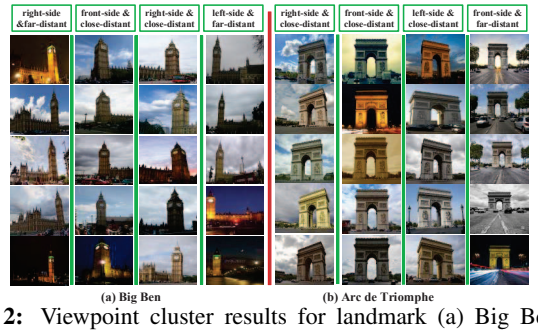

Fig. 2: Viewpoint cluster results for landmark (a) Big Ben and (b)Arc de Triomphe

and cloudy scenes. In other words, when an image is represented by Bag-of-Words, the landmark images with the same scene should share the same visual words, which are distinctive from others. Considering that the BoWs of an image can be regarded as the words of a document, we can utilize the topic model to discover these distinctive visual words for scenes from each viewpoint. Moreover, each viewpoint cluster is composed of all the scene categories from many different people like tourists, photographers, etc. Thus we design a new topic model, namely scene-viewpoint based theme model, including two layers: the parent layer and child layer. In the parent layer, we discover the so-called **Common Scene Theme** from the images of all the viewpoint clusters. The visual words contributing to the same common scene theme are probably shared by all viewpoint clusters but corresponding to the same scene. In the child layer, we discover the so-called **Scene-Viewpoint Theme** from the images of the same viewpoint cluster. Those visual words contributing to the same scene-viewpoint theme are specific to one certain viewpoint. Besides the common scene theme and scene-viewpoint one, there should be some non-discriminative and non-informative visual words, corresponding to the **Background Theme** shared by all the images. The visual words of all the images can be generated by these three kinds of themes.

In order to completely represent each scene theme using the visual words from three kinds of themes, we should align three kinds of themes through the same theme index z. When generating one image from a viewpoint cluster g, one firstly chooses the theme index z, and then draws visual words from three kinds of theme-word distributions, namely common scene theme-word distribution ϕ_z^C, scene-viewpoint theme-

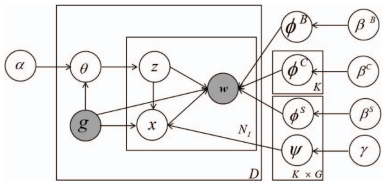

Fig. 3: The graphical representation of the theme modeling

word distribution $\phi^S_{g,z}$ and background theme-word distribution ϕ^B. The selection of the distributions is controlled by introducing the variable x. The details of generative process are as follows:

1. draw $\phi^B \sim Dir(\beta^B)$, $\phi^C_z \sim Dir(\beta^C)$, $\phi^S_{g,z} \sim Dir(\beta^S)$
2. draw a multinomial distribution $\psi_{g,z} \sim Dir(\gamma)$
3. for each image I
   (a) draw a theme mixture $\theta_I \sim Dir(\alpha)$
   (b) for the ith word in image I
      i. draw a theme $z_i \sim Mult(\theta_I)$
      ii. draw $x_i \sim Mult(\psi_{g,z_i})$
      iii. if $x_i = 0$, draw a word $w_i \sim Mult(\phi^B)$
      if $x_i = 1$, draw a word $w_i \sim Mult(\phi^C_{z_i})$
      if $x_i = 2$, draw a word $w_i \sim Mult(\phi^S_{g,z_i})$

The corresponding generative graph model can be represented in Fig. 3. Note that α, β^B, β^C,β^S and γ are hyper-parameters of the symmetric Dirichlet priors.

3.1. Approximate Inference

The goal of the inference in theme modeling is to estimate the new assignments **z** and **x** given the observed words **w**. We use Gibbs Sampling [15] for approximate inference in an iterative way, which alternates between estimating **z** and **x** as follows:

$$p(z_i|\mathbf{x}, \mathbf{z_{-i}}, \mathbf{w}, \Phi) \propto \begin{cases} (n_I^{z_i,-i} + \alpha_{g_I z_i}) \times \frac{n^{w_i,-i}_{G,x_i} + \beta^B}{n^{(\cdot),-i}_{G,x_i} + W\beta^B}, & x_i = 0 \\ (n_I^{z_i,-i} + \alpha_{g_I z_i}) \times \frac{n^{w_i,-i}_{z_i,x_i} + \beta^C}{n^{(\cdot),-i}_{z_i,x_i} + W\beta^C}, & x_i = 1 \\ (n_I^{z_i,-i} + \alpha_{g_I z_i}) \times \frac{n^{w_i,-i}_{g_I,z_i,x_i} + \beta^S}{n^{(\cdot),-i}_{g_I,z_i,x_i} + W\beta^S}, & x_i = 2 \end{cases} \quad (4)$$

$$p(x_i|\mathbf{x_{-i}}, \mathbf{z}, \mathbf{w}, \Phi) \propto \begin{cases} (n^{x_i,-i}_{g_I,z_i} + \gamma) \times \frac{n^{w_i,-i}_{G,x_i} + \beta^B}{n^{(\cdot),-i}_{G,x_i} + W\beta^B}, & x_i = 0 \\ (n^{x_i,-i}_{g_I,z_i} + \gamma) \times \frac{n^{w_i,-i}_{z_i,x_i} + \beta^C}{n^{(\cdot),-i}_{z_i,x_i} + W\beta^C}, & x_i = 1 \\ (n^{x_i,-i}_{g_I,z_i} + \gamma) \times \frac{n^{w_i,-i}_{g_I,z_i,x_i} + \beta^S}{n^{(\cdot),-i}_{g_I,z_i,x_i} + W\beta^S}, & x_i = 2 \end{cases} \quad (5)$$

where the superscript $-i$ denotes a counting variable that excludes the i-th word index in the corpus, and the superscript $(\cdot)$ denotes a counting variable that sums over all elements in the corresponding vector. $n_I^{z_i,-i}$ is the number of word tokens assigned to theme z_i in image I. $n^{x_i,-i}_{g_I,z_i}$ is the number of word tokens assigned to background ($x_i = 0$), common scene ($x_i = 1$) or scene-viewpoint ($x_i = 2$) theme z_i in collection g_I. $n^{w_i,-i}_{G,x_i}$, $n^{w_i,-i}_{z_i,x_i}$ and $n^{w_i,-i}_{g_I,z_i,x_i}$ is the number of times that word token w_i is assigned to all viewpoint clusters collection G, common scene theme z_i and scene-viewpoint theme z_i for viewpoint g respectively.

3.2. Summary generation

The landmark summary is generated by the discovered common scene themes and scene-viewpoint themes from all viewpoint clusters. For each scene from each viewpoint cluster, we select top-ranked images, where the images are sorted by counting the theme indicator variables of visual words $\mathbf{w}_{c,I}$ from the common scene theme and $\mathbf{w}_{s,I}$ from the scene-viewpoint theme:

$$p(z_k|\mathbf{w}_{c,I}, \mathbf{w}_{s,I}) = \frac{\sum_{i=1}^{n^v_{c,I}} \mathbf{I}(z^w_{c,I,i} = k) + \sum_{i=1}^{n^w_{s,I}} \mathbf{I}(z^w_{s,I,i} = k)}{n^w_{c,I} + n^w_{s,I}} \quad (6)$$

where $z^w_{c,I,i}$ and $z^w_{s,I,i}$are the theme assignments for the i-th common scene word and scene-viewpoint word for image I, and $n^w_{c,I}$ and $n^w_{s,I}$ denotes the number of visual words from common scene themes and scene-viewpoint theme in image I, respectively. $\mathbf{I}(\cdot)$ is an indicator function returning 1 if it is true and 0 otherwise.

4. EXPERIMENT

The experiments are performed on 10 landmarks, shown in Table 1. We crawl images from Flickr with the landmark name as the query. As we aim to summarize landmarks, we implement face detection using Face++ [16] to filter out images with faces. Table 1 details the statistics of resulted images.

Table 1: The statistics of our collected landmarks

Landmark	♯image	Landmark	♯image
Big Ben	13136	Notre Dame	8822
Statue of Liberty	7575	Arc de Triomphe	8487
Golden Gate Bridge	10745	Brandenburg Gate	7582
Tower Bridge	10814	Parthenon	7153
Eiffel Tower	12419	Basil Cathedral	8458

For the feature extraction, we extract 512-D GIST features [17] for viewpoint clustering. While for Scenic theme modeling, we first segment each image into regions using N-Cuts [18] and then extract 869-D global features including Texton Histograms and Color Histograms [19] for each region. 100,000 region features are sampled from all landmarks and are further quantized to constitute a codebook of 1024 visual words. In addition, the number of viewpoint clustering is empirically set to 20. In order to show images with more representative viewpoints, we select 5 top-ranked viewpoint clusters for scenic theme modeling. We report the best result with the number of scene themes $K = 8$, $\alpha = 1.0$, $\beta^B = \beta^C = \beta^S = 0.01$, $\gamma = 1.0$. The iteration number of Gibbs sampling is 2000.

To evaluate our method, we consider the following baselines for comparison:

- VC: Viewpoint Clustering.
- LDA: Latent Dirichlet Allocation [20].

For all methods, we select the same number of images for each landmark to ensure the fair comparison. Particulary, for

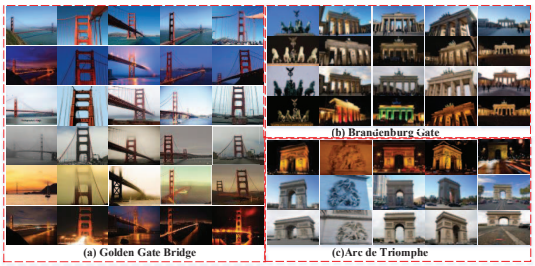

Fig. 4: Summarized Results for landmark (a) Golden Gate Bridge (b) Brandenburg Gate and (c) Arc de Triomphe. For each landmark, each row represents one scene and each column represents one viewpoint.

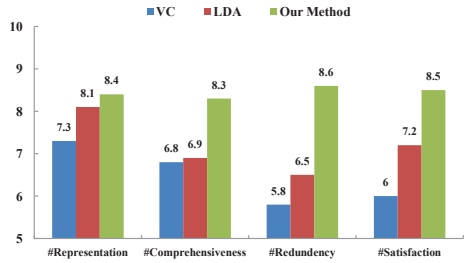

Fig. 5: User experience comparison between Our method and two baselines from four aspects.

our method, since the number of scene themes is 8 for each viewpoint cluster, the number of scene themes from all viewpoint clusters is $8 \times 5 = 40$. We select 5 images for each scene theme from each viewpoint cluster, where the images are sorted by Eqn.6; For VC, we select 5 top-ranked viewpoint clusters from Viewpoint Clustering, which is the same as our method, and then select 40 top-ranked images from each viewpoint cluster, where the images are ranked by Eqn.3. For LDA, we group the 5 top-ranked viewpoint clusters together and directly applied LDA to the whole collection. The number of themes is set the same as our method, that is 40. For each theme, we select 5 top-ranked images by $p(z|I)$.

We firstly conduct the quantitative evaluation of the proposed method. Particulary, we randomly select 4 landmarks and ask 20 users to label semantically meaningful themes for each landmark. Note that we consider one viewpoint cluster for VC as one theme. The statistics on the number of discovered themes are shown in Table 2. Compared with two baselines, which only focus on either the viewpoint or scene, our method is capable of differentiating among themes based on both viewpoints and scenes and thus discovers more meaningful themes.

Table 2: The statistics of our collected landmarks

Method	Big Ben	Eiffel Tower	Notre Dame	Tower Bridge	Mean
VC	5	5	5	5	5
LDA	22	24	26	21	23
Our Method	25	30	30	25	27

For landmark summarization, the first top-ranked image in each theme is selected to show the different viewpoints and scenes of the corresponding landmark. Fig.4 shows the visual summarization for three landmarks Golden Gate Bridge, Brandenburg Gate and Arc de Triomphe. The images in each row corresponds to one scene. For example, the second last row in Golden Gate Bridge (Fig. 4 (a)) is sunset, and the last row in Brandenburg Gate (Fig. 4 (b)) is night view with lights. The images in each column corresponds to a certain viewpoint. For example, the second column in Golden Gate Bridge is front-side and close-up view, and the last column of Arc de Triomphe (Fig. 4 (c)) is front-side and far-distant view.

To evaluate our method, we asked these 20 users to judge all the images on four aspects: representation, comprehensiveness, redundancy and satisfaction. Users were asked to score every landmark with 0 to 10 on these four aspects (10 is the highest score). The results shown in Fig.5 are the average values of all the users and landmarks for each method. We can see that our method receives significant gains in comprehensiveness, redundancy and satisfaction. The score on comprehensiveness of our method is the highest. This is because our method summarizes landmarks based on both viewpoint and scenes, while other two baselines only focus on either the viewpoint or scene. The score on redundancy of our method is higher than LDA, and LDA is higher than VC, while that on satisfaction of our method is higher than these two baselines. The score on representation of our method is comparable to LDA.

5. CONCLUSION

We have proposed a framework, which can summarize landmarks with diverse scenes and viewpoints. In the framework, we firstly group images into different clusters by the viewpoint and then discover distinctive scenes from each viewpoint cluster via scene-viewpoint based theme modeling. The experiments have shown the effectiveness of our method. In the future work, we plan to apply our method to more landmarks and evaluate its generalization ability. Furthermore, applications like travel recommendation will be designed using the proposed landmark summary framework.

Acknowledgement

This work is supported in part by National Basic Research Program of China (No. 2012CB316304), National Natural Science Foundation of China (No. 61225009, No. 61201374) and Beijing Natural Science Foundation (No. 4131004). This work is also supported by the Singapore National Research Foundation under its International Research Centre@Singapore Funding Initiative and administered by the IDM Programme Office.

6. REFERENCES

[1] Y. Zheng, M. Zhao, Y. Song, H. Adam, U. Buddemeier, A. Bissacco, F. Brucher, T.S. Chua, and H. Neven, "Tour the world: building a web-scale landmark recognition engine," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2009, pp. 1085–1092.

[2] B. Bao, B. Ni, Y. Mu, and S. Yan, "Efficient region-aware large graph construction towards scalable multi-label propagation," *Pattern Recognition*, vol. 44, no. 3, pp. 598–606, 2011.

[3] Y. Avrithis, Y. Kalantidis, G. Tolias, and E. Spyrou, "Retrieving landmark and non-landmark images from community photo collections," in *Proceedings of the international conference on Multimedia*. ACM, 2010, pp. 153–162.

[4] W. Min, C. Xu, M. Xu, X. Xiao, and B. Bao, "Mobile landmark search with 3d models," *IEEE Transactions on Multimedia*, vol. 16, no. 3, pp. 623–636, 2014.

[5] X. Xiao, C. Xu, J. Wang, and M. Xu, "Enhanced 3-d modeling for landmark image classification," *IEEE Transactions on Multimedia*, vol. 14, no. 4, pp. 1246–1258, 2012.

[6] R. Ji, Y. Gao, B. Zhong, H. Yao, and Q. Tian, "Mining flickr landmarks by modeling reconstruction sparsity," *ACM Transactions on Multimedia Computing, Communications, and Applications*, vol. 7, no. 1, pp. 31, 2011.

[7] R. Raguram, C. Wu, J. Frahm, and S. Lazebnik, "Modeling and recognition of landmark image collections using iconic scene graphs," *International journal of computer vision*, vol. 95, no. 3, pp. 213–239, 2011.

[8] I. Simon, N. Snavely, and S. Seitz, "Scene summarization for online image collections," in *IEEE 11th International Conference on Computer Vision*, 2007, pp. 1–8.

[9] L. Kennedy and M. Naaman, "Generating diverse and representative image search results for landmarks," in *Proceeding of the 17th international conference on World Wide Web*. ACM, 2008, pp. 297–306.

[10] Y. Zhao, Y. Zheng, X. Zhou, and T. Chua, "Generating representative views of landmarks via scenic theme detection," in *Advances in Multimedia Modeling*, pp. 392–402. Springer, 2011.

[11] R. Ji, H. Yao, W. Liu, X. Sun, and Q. Tian, "Task-dependent visual-codebook compression," *IEEE Transactions on Image Processing*, vol. 16, no. 3, pp. 623–636, 2014.

[12] B. Bao, G. Liu, R. Hong, S. Yan, and C. Xu, "General subspace learning with corrupted training data via graph embedding," *IEEE Transactions on Image Processing*, vol. 22, no. 11, pp. 4380–4393, 2013.

[13] X. Yao and X. Qian, "Visual summarization of landmarks via viewpoint modeling," in *IEEE 19th International Conference on Image Processing*, 2012, pp. 2873–2876.

[14] D. Verma and M. Meila, "A comparison of spectral clustering algorithms," *Technical Report, Department of CSE University of Washington Seattle,WAWA, 98195C2350*, 2005.

[15] T. L. Griffiths and M. Steyvers., "Finding scientific topics," *Proceedings of the National academy of Sciences of the United States of America*, vol. 101, no. Suppl 1, pp. 5228–5235, 2004.

[16] "Face++," http://cn.faceplusplus.com/uc/doc/home.

[17] A. Oliva and A. Torralba, "Modeling the shape of the scene: A holistic representation of the spatial envelope," *International journal of computer vision*, vol. 42, no. 3, pp. 145–175, 2001.

[18] J. Shi and J Malik, "Normalized cuts and image segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 8, pp. 888–905, 2000.

[19] Y.J. Lee and K.Grauman, "Object-graphs for context-aware category discovery," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2010, pp. 1–8.

[20] D. Blei, A. Ng, and M. Jordon, "Latent dirichlet allocation," *the Journal of machine Learning research*, vol. 3, pp. 993–1022, 2003.