Image and Video Representations based on Visual Dictionaries

Otávio A. B. Penatti[1], Eduardo Valle[1,2], Ricardo da S. Torres[1]
[1]RECOD Lab, Institute of Computing (IC)
[2]Dept. of Comp. Eng. and Ind. Aut. (DCA), School of Electrical and Comp. Eng. (FEEC)
University of Campinas (Unicamp), Campinas, SP - Brazil
penatti@ic.unicamp.br, dovalle@dca.fee.unicamp.br, rtorres@ic.unicamp.br

***Abstract*—The thesis explores three research topics involving the popular approach used for representing visual content: the visual dictionaries. The first topic concerns the generality of visual dictionaries: does a dictionary based on one dataset generalize to another dataset? Our findings create the opportunity to greatly alleviate the burden in generating dictionaries. The second topic is related to the importance of the spatial information of visual words in the image space for distinguishing types of scenes and objects. We propose an efficient and effective spatial pooling approach which presents promising results for image retrieval. And the third topic refers to the semantic information in the visual dictionary model. We claim that a bag-of-prototypes model, where the prototypes are visual words carrying semantics, is promising for improving image and video representations. Employing this model, we propose a semantically enriched dictionary based on scenes, which was effectively used for video geocoding. Defended in November 29th, 2012, the thesis has already generated 6 publications, including a best paper award. One of the proposed approaches has also obtained one of the best results in the Placing Task of MediaEval challenge in the last two years[1].**

***Keywords*-visual dictionaries; computer vision; image representation; video representation;**

I. Introduction and motivation

Representing images based only on their content has been challenging researchers and companies for decades. Many steps towards the objective of making a machine able to understand what it sees have been successful, but many others are still necessary. The thesis aims at contributing in smoothing the next steps in this direction.

The current advances in technology are changing the way how people live, specially considering the impact brought by the Internet and the image capturing devices. Recently, due to the increasing computational power of digital devices, people are getting in touch with systems based on powerful computer vision approaches. We can notice, for instance, the popularity of face recognition algorithms embedded into digital cameras and the trend of mobile applications like Google Goggles[2]. All those kinds of applications are dependent on representing image visual properties effectively.

The research community has developed several different methods for representing image visual properties. *Global descriptors* [1], [2], for example, have the advantage of being simple to compute but they encode few local properties of images. Thus, for object recognition and more precise applications, they can be less effective. *Local descriptors* [3] are powerful to encode local properties, but they are computationally more expensive, generate a variable number of feature vectors per image, and are very sensitive to variations in objects, limiting their use to some specific applications, like copy detection.

In the year of 2003, a method proposed by Sivic and Zisserman [4] introduced the idea of representing images in a similar fashion as representing text documents. As well as a text document is composed of a set of textual words, an image can be analyzed as a set of local appearances, also called visual words. A visual dictionary, often called visual codebook, is the set of local appearances available to describe image content. This approach is based on the use of local descriptors, however, by using the visual dictionary, we can compute statistics from the visual words present in the image, resulting in a single feature vector per image: the popular *bag of (visual) words* (BoW). Another advantage of visual dictionaries is that the description is more general, eliminating the problem of very precise representations generated by local descriptors and making the BoW representations useful in a wider range of applications. Figure 1 shows the potential of the representations based on visual dictionaries.

The process of generating and using the visual dictionary has several challenges and the thesis addresses some of them.

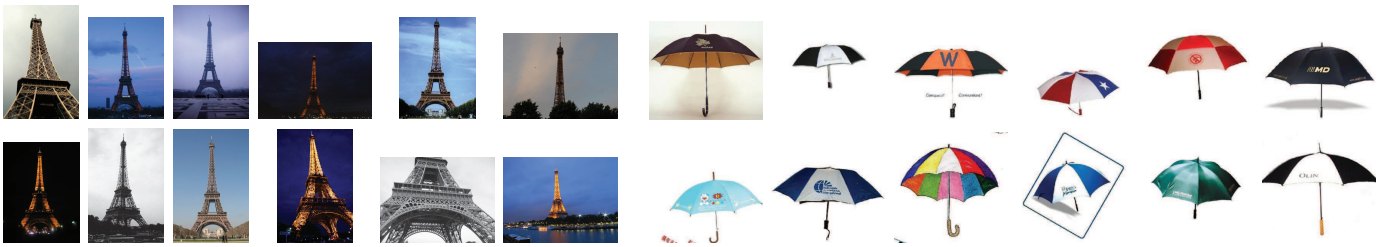
A. Hypotheses and research questions

The hypotheses analyzed in the thesis are the following:

- Visual dictionaries generalize well from one dataset to another, and from a subset of the classes to a whole dataset.
- The spatial information of visual words in the image space is important to distinguish types of scenes and objects.
- The use of semantically enriched dictionaries improves the quality of image and video representations.

In the following sections, we give the background for the statement of each hypothesis, show how we dealt with them and the contributions obtained.

[1]This work is related to a Ph.D. Thesis.
[2]www.google.com/mobile/goggles/ (as of April 26th, 2013).



(a) (b)

Fig. 1. Examples showing the potential of visual dictionaries. In (a), we show that even when the object of interest suffers transformations, like illumination, point of view, and scale, the representations remain similar. In (b), we show different instances of objects of the same type, which are considered similar using a visual dictionary representation. The examples shown are based on ranking the images which are represented by the proposed WSA descriptor (see Section III) in (a) Paris and (b) Caltech-101 datasets.

II. Hypothesis 1: Visual Dictionaries are Generalizable

Our first hypothesis under analysis in the thesis concerns the generality of visual dictionaries.

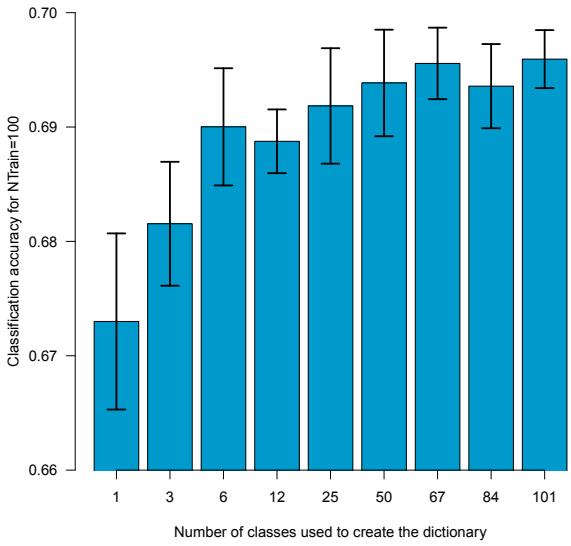
The first step to create a visual dictionary is the extraction of low-level features from images, which is usually performed by local descriptors. After extracting such features, the feature space is quantized in order to generate the visual dictionary [5]. A dictionary is usually generated based on a set of features extracted from images. We question if the set of images used really makes difference in the dictionary quality. How should we handle the cases in which the set of images is completely different from the images to be represented? In a Web-like environment, how should we deal with the fact that many images are constantly being inserted and removed from the dataset? The new images would still be well represented by previously created dictionaries?

We performed experiments on traditional datasets and on a large dataset of Web images to verify the impact of using different sources of information for creating visual dictionaries. We tested, for instance, if a dictionary created on a given dataset is good to represent images of another dataset. We also tested if a dictionary created on a small sample of classes is good to represent an entire collection. We could evaluate the feasibility of using visual dictionaries in scenarios where the entire dataset is not available for the dictionary construction as, for example, in large-scale dynamic datasets, like the Web.

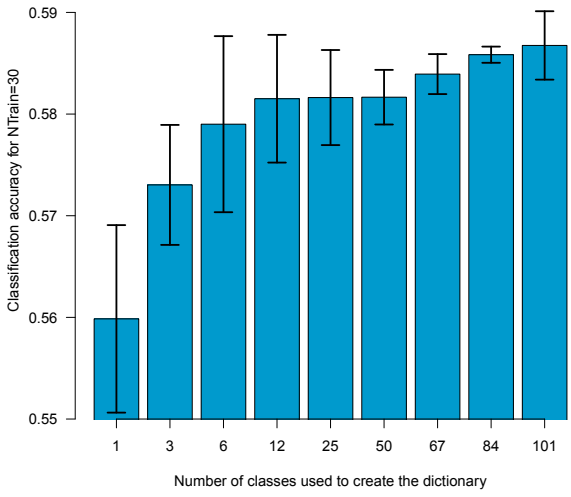
The experiments conducted show that dictionaries based on a subset of the collection, or even on an entirely different collection, may still provide good performance, on the condition that the selected sample is visually diverse. See, for instance, the results presented in Figure 2.

Therefore, we could confirm the hypothesis. Visual dictionaries generalize among datasets with similar characteristics, that is, similar datasets in terms of visual diversity may be used to generate good dictionaries for other datasets of the same kind.

Those findings open the opportunity to greatly alleviate the burden in generating the codebook, since, at least for general-purpose datasets, we show that the dictionaries do not have to take into account the entire collection, and may even be



(a) 15-Scenes



(b) Caltech-101

Fig. 2. Classification accuracy on (a) 15-Scenes and (b) Caltech-101 datasets using 9 different dictionaries created over a variable number of classes from Caltech-101. Although the results show some random fluctuation, it is clear that as soon as we have higher *visual* diversity, the accuracy reaches its asymptotic value, even if *semantically* (in terms of class diversity), the sample is still very poor.

based on another small collection of well-chosen visually diverse images. The contributions regarding this hypothesis are reported in a scientific paper which is under review in the *Image and Vision Computing* journal.

III. Hypothesis 2: Spatial Information of Visual Words is Important

The second hypothesis analyzed in the thesis refers to the importance of the spatial information of visual words in the image space.

The traditional pooling strategies [6], which compute the final bag of words (BoW) for an image, usually discard the positions of the visual words in the image space and considers only their activation levels in the feature space. However, by changing the spatial arrangement of image local patches we may also change image semantics. In the past, researchers faced the problem of having images with similar color histograms but different semantics. In the BoW representation, we migrate the problem from pixels to local patches.

Literature has a vast range of techniques [7]–[9] targeting the encoding of spatial information of visual words in the image space. The most popular approach is based on Spatial Pyramids [7]. Although they lead to large improvements on classification experiments, their huge vector is a problem in image retrieval applications. Many other approaches suffer from the same problem of generating large feature vectors [9] and others target specific applications [8].

A second contribution of the thesis is a pooling method that encodes the spatial arrangement of visual words in an image, called *Word Spatial Arrangement (WSA)*. WSA encodes the relative position of visual words in the image by splitting the image space using each point as the origin of a four-quadrant structure and counting the number of points in each quadrant, as exemplified in Figure 3. WSA increases the discriminating power of non-spatial pooling approaches keeping one of the BoW strengths, that is the general aspect of the representation. WSA has the benefits of generating more compact vectors than most of the spatial pooling methods. WSA is flexible enough to work both for image retrieval and classification and it works well in both hard and soft assignments. This flexibility is not supported by many of the existing spatial pooling methods.

We have performed experiments using four different image datasets analyzing the performance of WSA under different criteria. Experiments for image retrieval show that WSA outperforms the most popular approach to spatial pooling, the Spatial Pyramids (see Figure 4). We also provide an on-line interface to navigate through the results[3] [4].

Therefore, we could confirm the second hypothesis and also propose an efficient and effective solution for spatial pooling. The results in this topic were published in the *Iberoamerican Congress on Pattern Recognition* (CIARP) [10], in 2011, receiving the **best paper award**. An extension of that paper containing the results presented in the thesis is under review in the *Pattern Recognition* journal.

[3]www.recod.ic.unicamp.br/eva/view_images_base600.php (as of April 26th, 2013).

[4]www.recod.ic.unicamp.br/eva/view_images_paris.php (as of April 26th, 2013).

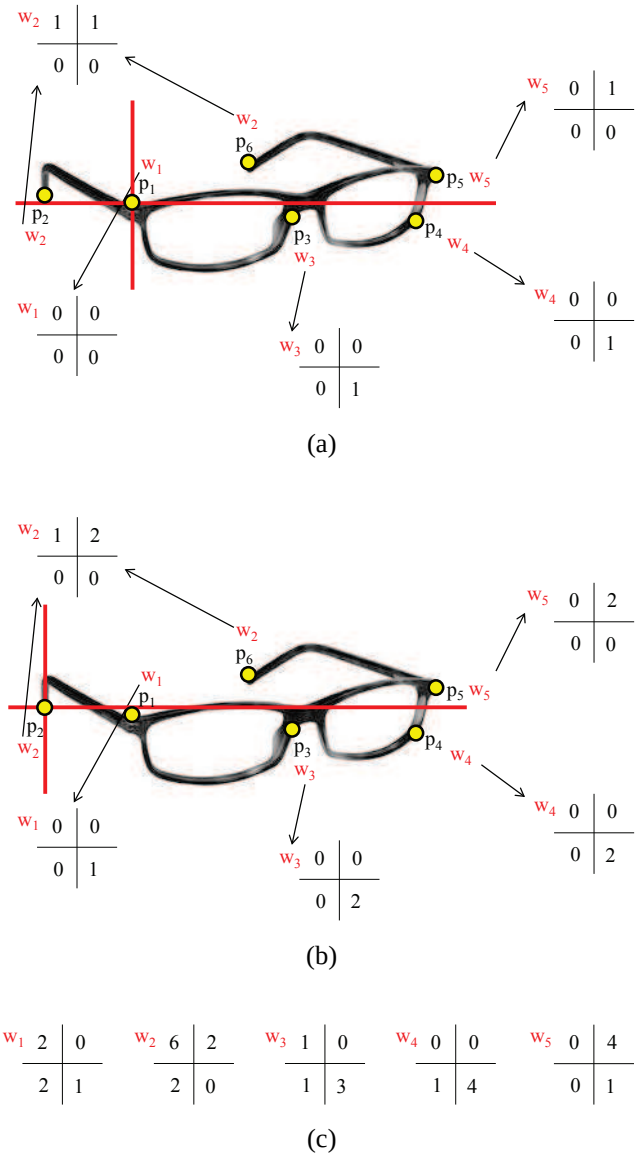


Fig. 3. Example of WSA's partitioning and counting. The small circles are the detected points, tagged with their associated visual words (w_i's). We start in (a), putting the quadrant's origin at p_1 and counting in the visual word associated with each other point, where the point is in relation to p_1. On the second step (b) the quadrant is at p_2; we increment again the counters of the words associated with each other point in the position corresponding to their position in relation to p_2. We proceeded until the quadrant has visited every point in the image. Final counter values for this example are shown in (c).

IV. Hypothesis 3: Semantically Enriched Dictionaries Improve the Quality of the Representations

The third hypothesis under analysis in the thesis is related to the semantic information in the visual dictionary model.

An important aspect of visual dictionaries based on local features lies in the fact that visual words carry little or no

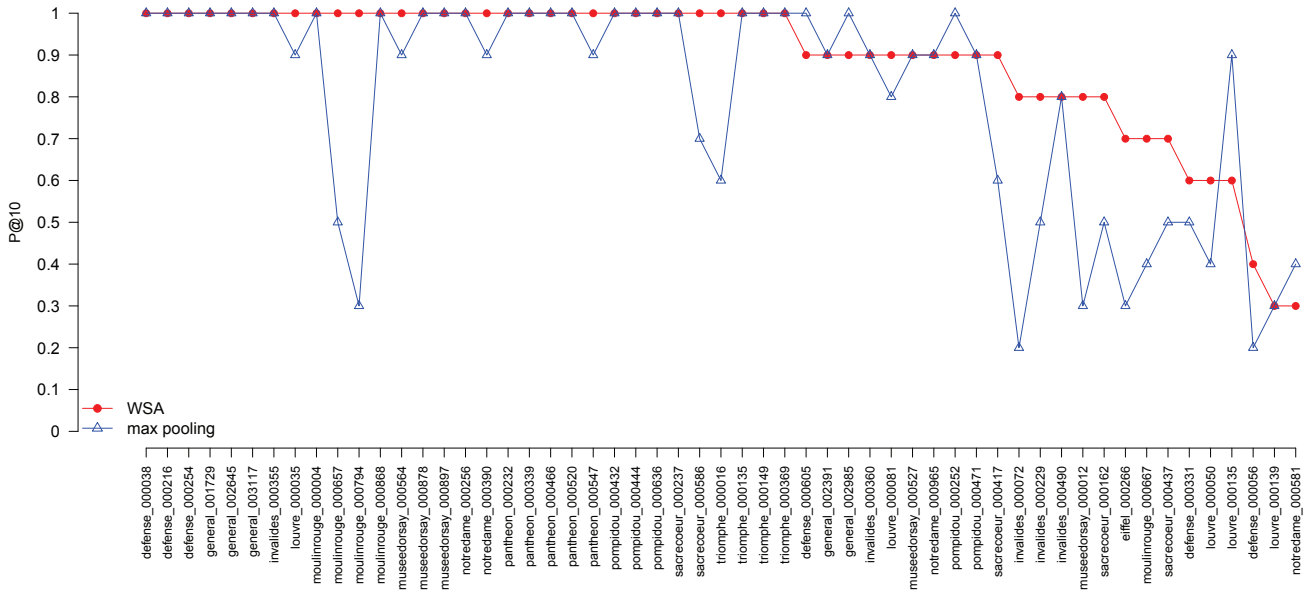


Fig. 4. S-curves showing that the proposed WSA representation (red line) outperforms the best max pooling configuration (blue line) in the Paris dataset. The S-curves highlight the performance (vertical axis) of the two methods for each of the queries (horizontal axis). Performance measure: P@10.

semantics [11]. Therefore, the term *dictionary* is somewhat misleading, because their words have no meaning for humans. However, the representations based on visual dictionaries are powerful. Thus, when we move from the low-level feature space to the mid-level (BoW) space, we obtain a semantic separability that makes it possible to distinguish different types of scenes and objects. What would happen if we move one step further by using a dictionary where the visual words have more semantic information?

We first performed an analysis on the semantic information involved in the feature spaces comprised by the visual dictionary model. We analyzed the separability between distance distributions considering different semantic classes of points or objects. In the low-level feature space, although we could expect that appearances carry semantics, we show that there is no semantic separability between distance distributions. In the mid-level space, despite the good results of BoW representations in the literature, we show that the semantic separability between distance distributions in this space is still small.

This motivates the creation of a new feature space with more semantics. In this direction, we have worked on a *bag-of-prototypes* model, according to which the prototypes are elements containing more semantic information. By having visual words (prototypes) carrying semantics, we can compute a bag-of-prototypes representation which has an interesting property: each dimension has semantics by itself. Therefore, the feature space spanned by such model has the property of having one dimension for each semantic concept.

We employ this model proposing a new video representation based on a dictionary of scenes. Figure 5 compares a dictionary of scenes with a traditional dictionary based on local

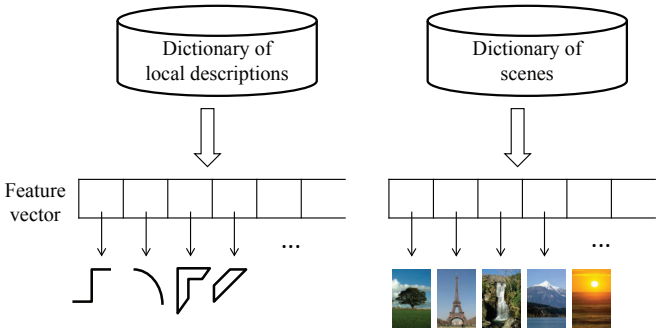


Fig. 5. Comparison between the proposed dictionary of scenes with a dictionary based on local descriptions. We can notice that the representation based on the local dictionary relies on elements without clear semantics, like small corners and edges, while, the representation based on the dictionary of scenes carries more semantics. In addition, the feature space for the dictionary of scenes has semantics in each dimension independently.

patches. The proposed representation, named *bag of scenes*, was evaluated in the context of video geocoding, which is the task of assigning a geographic location to videos. In this context, the bag-of-scenes vector works like a place activation vector, being informative for the geocoding task. Figure 6 shows how to create a dictionary of scenes and then how to create a bag of scenes.

We performed experiments under the Placing Task of the MediaEval 2011 challenge [12]. In this task, participants were required to assign a geographical location to a set of videos and the results were based on how far the videos were tagged from their correct locations. As presented in Table I, our results show that the proposed bag-of-scenes model is effective for video geocoding, being more precise than most of the visual

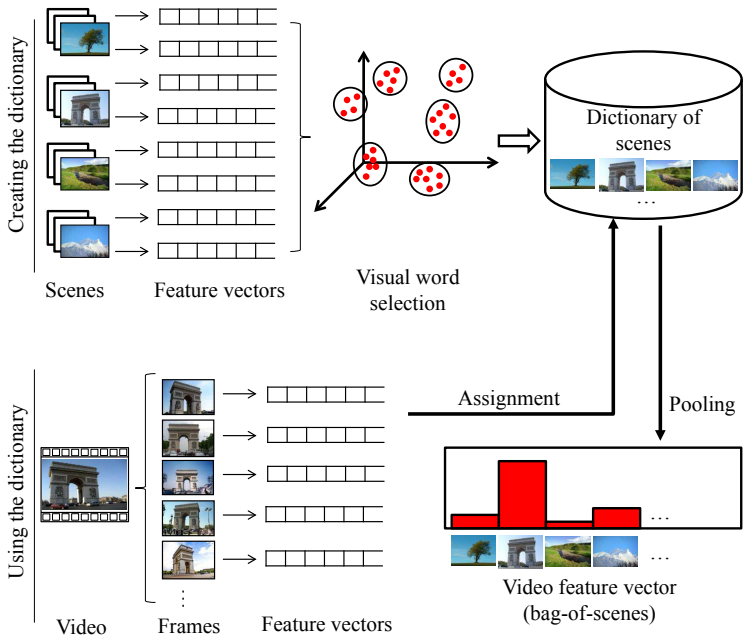


Fig. 6. Schema for generating and using a dictionary of scenes. The dictionary is created based on a given collection of scenes, which may come from an image dataset or from video frames. After representing each image with any kind of feature vector, some of them are selected to compose the dictionary. Given an input video to be represented, its frames are assigned to one or more of the scenes in the dictionary. A pooling strategy is then applied to generate the video feature vector (*bag of scenes*).

geocoding methods presented at the Placing Task of 2011.

The accomplished results, thus, confirm the hypothesis. The proposed bag-of-scenes model was published in the *ACM International Conference on Multimedia Retrieval* (ICMR), in 2012 [13]. We also participated in the Placing Task of MediaEval 2012 [14], achieving the best results with the use of only visual information.

V. Publications

The thesis has already generated a set of 6 publications and 2 working notes. As mentioned in the previous sections, some additional works are under review in international journals. The publications are the following:

- *Encoding spatial arrangement of visual words* [10], O. A. B. Penatti, E. Valle, and R. da S. Torres, in the Iberoamerican Congress on Pattern Recognition (CIARP), 2011, receiving the **best paper award**. An extension of this work is under review in the Pattern Recognition journal.
- *A Visual Approach for Video Geocoding using Bag-of-Scenes* [13], O. A. B. Penatti, L. T. Li, J. Almeida, and R. da S. Torres, in the International Conference on Multimedia Retrieval (ICMR), 2012.
- *Improving Texture Description in Remote Sensing Image Multi-Scale Classification Tasks By Using Visual Words* [15], J. A. dos Santos, O. A. B. Penatti, R. da S. Torres, P-H. Gosselin, S. Philipp-Foliguet, and A. X. Falcão, in the International Conference on Pattern Recognition (ICPR), 2012.
- *Multimedia Multimodal Geocoding* [16], L. T. Li, D. C. G. Pedronette, J. Almeida, O. A. B. Penatti, R. T. Calumby, and R. da S Torres, in the International Conference on Advances in Geographic Information Systems (ACM SIGSPATIAL GIS), 2012.
- *UNICAMP-UFMG at MediaEval 2012: Genre Tagging Task* [17], J. Almeida, T. Salles, E. F. Martins, O. A. B. Penatti, R. da S. Torres, M. A. Gonçalves, and J. M. Almeida, in the Working Notes of the MediaEval Workshop, 2012.
- *A Multimodal Approach for Video Geocoding* [14], L. T. Li, J. Almeida, D. C. G. Pedronette, O. A. B. Penatti, and R. da S. Torres, in the Working Notes of the MediaEval Workshop, 2012.
- *Remote Sensing Image Representation based on Hierarchical Histogram Propagation*, J. A. dos Santos, O. A. B. Penatti, R. da S. Torres, P-H. Gosselin, S. Philipp-Foliguet, and A. X. Falcão, in the International Geoscience and Remote Sensing Symposium (IGARSS), *accepted for publication*, 2013.
- *Domain-specific Image Geocoding: A Case Study on Virginia Tech Building Photos*, L. T. Li, O. A. B. Penatti, E. A. Fox, and R. da S. Torres, in the Joint Conference on Digital Libraries (JCDL), *accepted for publication*, 2013.

VI. Conclusions

Making digital visual information understandable by computers is a challenge that motivates the research described in the thesis. One of the main elements to make this possible is the effective representation of the visual content. In the thesis, we focus on representations based on visual dictionaries. Visual dictionaries lie in the idea of describing visual content

TABLE I

COMPARING OUR BAG OF SCENES (2 RIGHTMOST COLUMNS) WITH THE OTHER TEAMS WHO PARTICIPATED IN THE PLACING TASK 2011 AND WORKED ONLY WITH VISUAL INFORMATION. THE TABLE SHOWS THE NUMBER OF TEST VIDEOS GEOCODED AT DIFFERENT DISTANCES FROM THE CORRECT VIDEO LOCATION. WE CAN HIGHLIGHT, FOR INSTANCE, THAT THE BAG OF SCENES IS EFFECTIVE FOR THE MOST PRECISE MEASURE (1KM).

Radius	**Other teams**				**Bag of Scenes**	
(km)	UGENT	UNICAMP	ICSI	WISTUD	Random_{5000}	Guided_{4000}
1	2	11	5	0	6	16
10	6	60	16	5	32	49
100	49	145	67	-	95	104
1 000	624	650	598	583	610	661
10 000	4 332	4 248	4 234	-	4 353	4 365

as describing text documents [4]. This model is successful for visual recognition, however, there are several challenges on how to make the dictionary based representations better or even on how to generate better dictionaries.

The thesis analyzed three hypotheses regarding the visual dictionary model. The first one concerns the generality of dictionaries. As dictionaries are result of a feature space quantization, which is usually performed on features extract from a set of images, we question if the set of images really makes difference in the dictionary quality. We experimentally show that by creating the dictionary on a set of images that is not related to the images to be represented, or even by creating a dictionary on a small sample of a dataset, we can still have a good dictionary. Those findings open the oportunity to greatly alleviate the cost for dictionary creation.

The second hypothesis is related to the importance of the spatial information of visual words in the image space. We propose a new spatial pooling method, called WSA, which is flexible to work for applications of both image retrieval and classification. The proposed WSA method is more compact than most of the spatial pooling methods and presented better results than the Spatial Pyramids for image retrieval. The paper proposing WSA received the *best paper award* in an international conference [10].

The third hypothesis under analysis in the thesis is related to the semantic information involved in the visual dictionary model. We worked on a *bag-of-prototypes* model on which the prototypes are visual words carrying semantic information. We proposed a new video representation, called bag of scenes [13], which employs that model. We evaluated the bag-of-scenes approach in the context of the Placing Task of the MediaEval challenge. The proposed approach presented one of the best results for visual-based geocoding in the last two years of the challenge [13], [14].

ACKNOWLEDGMENT

Authors are grateful to Fapesp (grant number 2009/10554-8), CAPES, CNPq, AMD, Microsoft Research, and Samsung for the financial support and infrastructure.

REFERENCES

[1] A. W. M. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain, "Content-based image retrieval at the end of the early years," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 12, pp. 1349–1380, 2000.

[2] O. A. B. Penatti, E. Valle, and R. da S. Torres, "Comparative study of global color and texture descriptors for web image retrieval," *Journal of Visual Communication and Image Representation*, vol. 23, no. 2, pp. 359–380, 2012.

[3] K. Mikolajczyk and C. Schmid, "A performance evaluation of local descriptors," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 10, pp. 1615–1630, 2005.

[4] J. Sivic and A. Zisserman, "Video google: a text retrieval approach to object matching in videos," in *International Conference on Computer Vision*, vol. 2, 2003, pp. 1470–1477.

[5] V. Viitaniemi and J. Laaksonen, "Experiments on selection of codebooks for local image feature histograms," in *International Conference on Visual Information Systems: Web-Based Visual Information Search and Management*, 2008, pp. 126–137.

[6] Y.-L. Boureau, F. Bach, Y. LeCun, and J. Ponce, "Learning mid-level features for recognition," in *Conference on Computer Vision and Pattern Recognition*, 2010, pp. 2559–2566.

[7] S. Lazebnik, C. Schmid, and J. Ponce, "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories," in *Conference on Computer Vision and Pattern Recognition*, vol. 2, 2006, pp. 2169–2178.

[8] J. Feng, B. Ni, Q. Tian, and S. Yan, "Geometric lp-norm feature pooling for image classification," in *Conference on Computer Vision and Pattern Recognition*, 2011, pp. 2609–2704.

[9] Y. Cao, C. Wang, Z. Li, L. Zhang, and L. Zhang, "Spatial–bag–of–features," in *Conference on Computer Vision and Pattern Recognition*, 2010, pp. 3352–3359.

[10] O. A. B. Penatti, E. Valle, and R. d. S. Torres, "Encoding spatial arrangement of visual words," in *Iberoamerican Congress on Pattern Recognition*, vol. 7042, 2011, pp. 240–247.

[11] L.-J. Li, H. Su, E. P. Xing, and L. Fei-Fei, "Object bank: A high-level image representation for scene classification and semantic feature sparsification," in *Neural Information Processing Systems*, 2010.

[12] A. Rae, V. Murdock, P. Serdyukov, and P. Kelm, "Working notes for the placing task at MediaEval," in *Working Notes Proceedings MediaEval Workshop*, vol. 807, 2011.

[13] O. A. B. Penatti, L. T. Li, J. Almeida, and R. da S. Torres, "A Visual Approach for Video Geocoding using Bag-of-Scenes," in *International Conference on Multimedia Retrieval*, 2012, pp. 53:1–53:8.

[14] L. T. Li, J. Almeida, D. C. G. Pedronette, O. A. B. Penatti, and R. da S. Torres, "A multimodal approach for video geocoding," in *Working Notes Proceedings MediaEval Workshop*, vol. 927, 2012.

[15] J. A. dos Santos, O. A. B. Penatti, R. da S. Torres, P.-H. Gosselin, S. Philipp-Foliguet, and A. X. Falcão, "Improving texture description in remote sensing image multi-scale classification tasks by using visual words," in *International Conference on Pattern Recognition*, 2012, pp. 3090–3093.

[16] L. T. Li, D. C. G. Pedronette, J. Almeida, O. A. B. Penatti, R. T. Calumby, and R. da S. Torres, "Multimedia Multimodal Geocoding," in *ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, 2012, pp. 474–477.

[17] J. Almeida, T. Salles, E. F. Martins, O. A. B. Penatti, R. da S. Torres, M. A. Gonçalves, and J. M. Almeida, "UNICAMP-UFMG at mediaeval 2012: Genre tagging task," in *Working Notes Proceedings MediaEval Workshop*, vol. 927, 2012.