Combining Holistic Descriptors for Scene Classification

Kelly Assis de Souza Gazolli and Evandro Ottoni Teatini Salles
Universidade Federal do Espírito, Vitória, ES, Brazil

Keywords: Visual Descriptor, Scene Classification, Gist Descriptor, Contextual Information, Census Transform, Holistic Approach.

Abstract: Scene classification is an important issue in the field of computer vision. To face this problem we explore in this paper a combination of Holistic Descriptors to scene categorization task. Therefore, we first describe the Contextual Mean Census Transform (CMCT), an image descriptor that combines distribution of local structures with contextual information. CMCT is a holistic descriptor based on CENTRIST and, as CENTRIST, encodes the structural properties within an image and suppresses detailed textural information. Second, we present the GistCMTC, a combination of Contextual Mean Census Transform descriptor with Gist in order to generate a new holistic descriptor representing scenes more accurately. Experimental results on four used datasets demonstrate that the proposed methods could achieve competitive performance against previous methods.

1 INTRODUCTION

Scenes classification is widely applied in many domains, such as, image retrieval and travel navigation. However, this task is quite challenging because there are a great number of possible classes. Also, some scenes are ambiguous and indeed, even human beings can be unsure about these classifications. In addition, the variation in illumination and scale could be daunting.

In the literature, several methods have been proposed to perform scene classification tasks. Oliva and Torralba (Oliva and Torralba, 2001) proposed a formal approach to build the ”gist” of the scene and provided a statistical summary of the spatial layout properties (naturalness, openness, expansion, depth, roughness, complexity, ruggedness, symmetry) of the scene.

Another popular approach is the local features (Grauman and Darrell, 2005) (Fei-Fei and Perona, 2005). In this method the image is divided into patches or regions on which individual features are computed. The collection of these local descriptors shapes the final representation. In this sense, Scalar Invariant Feature Transform (SIFT) (Lowe, 1999) became a very popular local descriptor. This approach transforms an image into a large collection of local feature vectors, each of which is invariant to translation, scaling and rotation and partially invariant to illumination changes.

The bag-of-features approach represents an image as an orderless collection of local features. This method models an image as an occurrence histogram of visual words that are local descriptors of regions or patches in the image (Wei Liu and Gabbouj, 2012). Many variants of this model have been proposed. Lazebnik et al. (Lazebnik et al., 2006) proposed a spatial pyramid, a technique which works by partitioning the image into increasingly fine sub-regions. Qin and Yung (Qin and Yung, 2010) proposed a method based on contextual visual words, in which the contextual information from neighbor region and the regions from coarser scales are included. Despite good results, this approach has some disadvantages. The codebook, i.e., the set of visual words, should be large enough so that each image could be properly represented by the histogram, thus, the codebook size depends on the dataset. Furthermore, the codebook-building process is often computationally intensive, which limits efficiency of its application (Wei Liu and Gabbouj, 2012).

Recently, Wu and Rehg (Wu and Rehg, 2011) proposed CENTRIST (Census Transform Histogram), a holistic representation that captures structural properties, rough geometry and generalizability by modeling distribution of local structures. CENTRIST is easy to implement, has nearly no parameter to tune, and is invariant to illumination.

In this paper, first, using The Contextual Mean Census Transform (CMCT), we show that combining

Gazolli K. and Salles E..
Combining Holistic Descriptors for Scene Classification.
DOI: 10.5220/0004286103150320
In *Proceedings of the International Conference on Computer Vision Theory and Applications* (VISAPP-2013), pages 315-320
ISBN: 978-989-8565-47-1

contextual features and local structures can help differentiate local structures that are similar but have considerable difference in its neighborhood. The Contextual Mean Census Transform is a holistic visual descriptor that captures structural properties, by modeling distribution of local structures, and adds contextual information, by modeling the distribution of structures formed by neighbor local structures. Then, we propose GistCMTC, a combination of CMTC contextual descriptor with Gist holistic approach. Therefore, by combining these two techniques we intend enhancing the descriptor quality by providing different types of information.

2 A CONTEXTUAL DESCRIPTOR

In this section, we first present the Modified Census Transform in which the proposed technique is based. Then we extend the concept to the Contextual Mean Census Transform (CMCT) descriptor.

2.1 Modified Census Transform

The Modified Census Transform (MCT) (Fröba and Ernst, 2004) is a nonparametric local transform based on Census Transform (Zabih and Woodfill, 1994). This technique was proposed with the aim of overcome some weakness of Census Transform. The Modified Census Transform, $\Gamma(x)$ is computed in the following manner. A 3 x 3 window of pixels is considered and the mean $\overline{I}(x)$ of the pixels is computed. Every pixel in the 3 x 3 window is then compared with $\overline{I}(x)$. If the pixel is bigger than or equal to $\overline{I}(x)$, a bit 1 is set in the corresponding location. Otherwise, a bit 0 is set, as follows

$$\Gamma(x) = \bigotimes_{y \in \mathcal{N}'(x)} \zeta(I(y), \overline{I}(x)), \quad \zeta(m,n) = \begin{cases} 1, & m \geq n \\ 0, & otherwise \end{cases} \tag{1}$$

where $\bigotimes$ represents concatenation operation, $\overline{I}(x)$ is the mean of the intensity values in the 3 x 3 window of pixels centered at x, $I(y)$ is the gray value of the pixel at y position and $\mathcal{N}'(x)$ is a local spatial neighborhood of the pixel at x. In the Modified Census Transform technique, 9 bits are generated and converted to a decimal number in [0, 511], namely, here, MCT.

In this work, we adopt Modified Census Transform. However, we do not compare the mean $\overline{I}(x)$ with the center pixel. Thus, we generate 8 bits, instead of 9, which are converted to a decimal number in [0, 255], i.e., we generate a smaller descriptor. Moreover, we adopt $\zeta(m,n) = 1$, if $m > n$. In order to differentiate Modified Census Transform with 9 bits from Modified Census Transform with 8 bits and with $\zeta(m,n) = 1$, if $m > n$, we refer to this last as MCT(8 bits). Finally, a 256 bins histogram of MCT(8 bits) values for an image is used as a visual descriptor.

2.2 CMCT - Contextual Mean Census Transform

The Contextual Mean Census Transform (CMCT) integrates contextual information with local structures information for differentiating regions that have similar structures, but have significant difference in their neighborhood. For accomplishing this task, this approach takes into consideration information of neighborhood windows in the MCT(8 bits) computation, by creating a new local structure from the local structure of the window and from the local structures of its neighboring windows. We believe these additional information can improve the image representation. We call these informations coming from outside windows by *context*.

The steps for the Contextual Mean Census Transform (CMCT) generation are as follow. First, MCT(8 bits) is computed for all pixels. Then, a histogram of MCT(8 bits) is obtained. A new image is created in which the original image pixels are replaced by the correspondent MCT(8 bits) values. In the sequel, the MCT(8 bits) is computed on the new images pixels and a new histogram is generated. Then, the CMCT(8 bits) histogram for the original image histogram and the the CMCT(8 bits) histogram for the new image are concatenated, generating a new descriptor. The whole process is schematized in Figure 1.

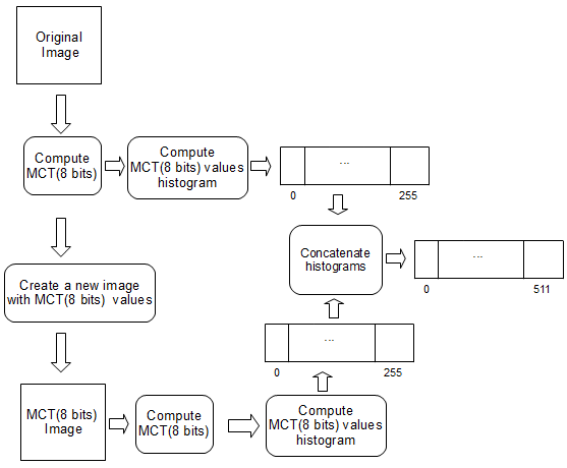

Figure 1: Contextual descriptor extraction process.

The MCT(8 bits) maps a 3 x 3 image window to one of 256 possible values. Therefore, the MCT(8

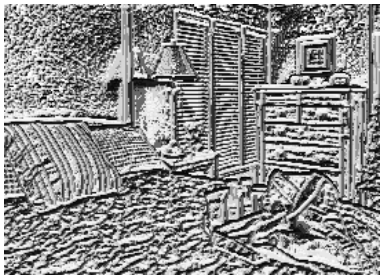

Figure 2: An example of an image from 15-category dataset with the gray values replaced by MCT(8 bits).

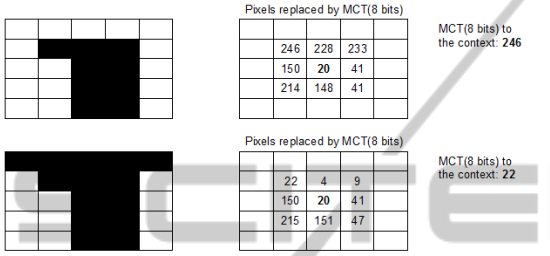

Figure 3: Two different windows with the same structure in the center (left). The pixels replaced by the values generated by MCT(8 bits) for each 3x3 window (center). The value of MCT(8 bits) to the new window (right).

bits) value acts as an index to different local structures. When pixel value is replaced by the MCT(8 bits) value, as illustrated in Figure 2, the information type changes from intensity to local structure index. So, when the Modified Census Transform is applied on the new image, it performs comparisons between local structures in the neighborhood, obtaining a different information type, i.e., the relationship between local structures. Figure 3 shows two examples of different 5 x 5 windows which has the same 3 x 3 window in the center. Although the windows have a similar center structure, they generate a different set of MCT(8 bits) values and, therefore, calculating the MCT(8 bits) in the new images generate different values for each example.

This technique does not represent all the possible local structures in a 5 x 5 window, which would require a 2^{25} size descriptor , but since the MCT(8 bits) for the original image and for the new image are computed, it is possible representing 65,536 (256 x 256) different structures using a 512 size descriptor. In this way, it is possible identifying a greater number of local structures and more accurately define their distribution in each class, improving scene classification results.

3 HOLISTIC APPROACHES

In this section we propose GistCMCT. This new descriptor is a combination of two holistic approaches: Gist (Oliva and Torralba, 2001) and CMCT.

3.1 Gist

Under the assumption that is not necessary identify the objects that make up a scene to identify the scene, Oliva and Torralba (Oliva and Torralba, 2001) proposed a holistic approach to build the ”gist” of the scene using low-dimensional representation of a set of global image properties such as naturalness, openness, roughness, expansion, and ruggedness. In this approach stable spatial structures within images that reflect functionality of the location are captured (Oliva and Torralba, 2001) rather than detailed information about objects. As Gist is a holistic and low-dimensional representation of the structure of a scene, it does not require explicit segmentation of image and objects. Therefore, this method requires very low computational resources (Oliva and Torralba, 2006).

In (Zabih and Woodfill, 1994) it was showed that Gist has a good performance when classifying outdoor scenes, however, when indoor scenes are added the Gist accuracy becomes worse.

3.2 GistCMCT

Although the techniques presented in section 2 are holistic, they differ from Gist. While Gist represents the shape of the scene by computing stable spatial structures within images that reflect functionality of the location, CMCT summarizes local shape information.

Gist has a certain weakness in recognizing indoor scenes, but is quite efficient in the recognition of outdoor scenes. CMCT, as the CENTRIST, represents structural properties through the distribution of local structures (for example, the percentages of local structures that are local horizontal edge) (Wu and Rehg, 2011) which helps in the classification of man-made environments, including, indoor environments. A vector composed of Gist and CMCT descriptors will gather their qualities, and will get, therefore, a better performance in classifying scenes.

4 EXPERIMENTS

In this section, we investigate the effectiveness of our representations and compare them with existing works.

4.1 Datasets and Setup

Our descriptors has been tested on four data sets: 8-category scenes provided by Oliva and Torralba (Oliva and Torralba, 2001), 15-category dataset (Lazebnik et al., 2006), 8-class sports event (Li and Fei-Fei, 2007) and 67-class indoor scene recognition (Quattoni and Torralba, 2009).

In the experiment, each category in a data set is split randomly into a training set and a test set. The random splitting is repeated 5 times, and the average accuracy is reported, as adopted by (Wu and Rehg, 2011). All color images were converted to gray scale. For training and classification, we adopted SVM (Support Vector Machine), a pattern classifier introduced by (Vapnik, 1998). We used the libSVM (Chang and Lin, 2011) package modified by (Wu and Rehg, 2009).

The *log* frequency weighting was applied in the histogram values. The *log* frequency weighting is a technique used in Information Retrieval (Salton and McGill, 1983) whereas relevance does not increase proportionally with term frequency. The *log* frequency weight of term t in a document d ($W_{t,d}$) is

$$W_{t,d} = \begin{cases} 1 + log(tf_{t,d}), & tf_{t,d} > 0, \\ 0, & otherwise \end{cases} \quad (2)$$

where $tf_{t,d}$ is number of occurrences of term t in a document d.

4.2 Experiments with CMCT

We first present the results from CMCT. For all experiments performed in this section, we employed linear kernel SVM to accomplish scene classification. Since the goal is to compare the efficiency of the descriptors, only the results of the experiments in which the images were not partitioned into increasingly fine subregions are considered. In such a case, it is used images without spatial representation and, therefore, with levels number equal to zero. CMCT was implemented using C++ and OpenCV.

4.2.1 15-Category Dataset

In this dataset an amount of 100 images in each category are used for training and the remaining images constitute the testing set, as in previous researches. When using CMCT, we achieve 76.87 $\pm$ 0.58% accuracy in this dataset. Figure 4 presents the confusion matrix from one run on 15-class scene dataset. We observe that the biggest confusion happens between bedroom and livingroom, which have similar elements. Humans may confuse them due to the small inter-class variation.

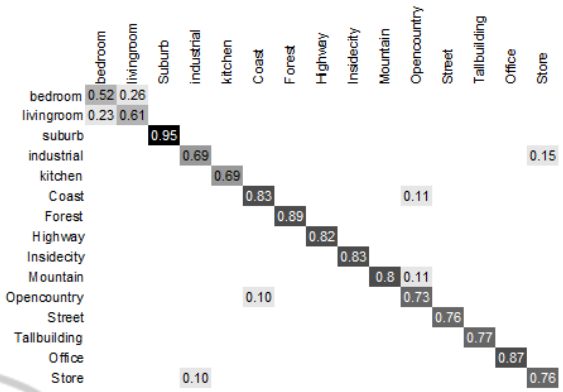

Figure 4: Confusion matrix from one run for 15-class scene recognition experiment using CMCT descriptor.

Table 1: Experimental results for 15-categories dataset.

Method	Accuracy(%)
CENTRIST without PCA	73.29 $\pm$ 0.96
SPM (16 channel weak features)	66.80 $\pm$ 0.6
SPM (SIFT 400 clusters)	74.80 $\pm$ 0.3
Gist	73.28 $\pm$ 0.67
RCVW	74.5
MCT(8 bits) Histogram	73.71 $\pm$ 0.30
CMCT	**76.87 $\pm$ 0.58**

Table 1 compares the classification performance of the proposed method on 15-category dataset to the following methods existing in literature: CENTRIST(Wu and Rehg, 2011), SPM (Lazebnik et al., 2006) with 0 levels, Gist (Oliva and Torralba, 2001), Region Contextual Visual Words (RCVW) (Liu et al., 2011) and Modified Centrist Transform (MCT) histogram with 8 bits (Fröba and Ernst, 2004), with CMCT.

In (Lazebnik et al., 2006), low level features were divided into weak features, computed from local 3 x 3 neighborhoods, and strong features, SIFT features computed from 16 x 16 image patches. CMCT outperforms the weak features and the strong features (SIFT 400 cluster centers). CMCT also outperforms CENTRIST, since the CMCT provides more information about the local image structures than CENTRIST. In (Liu et al., 2011), a method for scene categorization by integrating region contextual information into the bag-of-words approach is used. CMCT also overcomes this method. By comparing MCT(8 bits) histogram and CMCT, one can see that the addition of contextual information improves performance, since using only MCT(8 bits) histogram we achieve 73.71 $\pm$ 0.30%.

4.2.2 8-Category Dataset

In this dataset an amount of 100 images in each ca-

tegory are used for training and the remaining images constitute the testing set. In the 8-category scene class CMCT achieves 79.91 ± 0.99% accuracy.

Table 2 shows experimental results for 8-category dataset. Using Gist descriptor the recognition accuracy is 82.60 ± 0.86%, which is greater than the results achieved by the CMCT. However, on the 15-category dataset which adds several indoor categories, the accuracy using Gist dropped to 73.28 ± 0.67%, which is lower than CMCT accuracy. As in 15-category dataset, CMCT outperforms MCT(8 bits) histogram and CENTRIST.

Table 2: Experimental results for 8 scene categories dataset.

Method	Accuracy (%)
Gist	**82.60 ± 0.86**
CENTRIST (0 levels)	76.49 ± 0.84
MCT(8 bits) histogram	77.07 ± 0.68
CMCT	**79.91 ± 0.99**

4.2.3 8-Class Sports Event

Following (Li and Fei-Fei, 2007), in this dataset, we use 70 images per class for training and 60 for testing. CMCT achieves, in this dataset, 67.41 ± 1.10%, while CENTRIST (0 levels) achieves 63.91 ± 2.44%. In (Li and Fei-Fei, 2007), in which the event classification is a result of scene environment classification and object categorization, the accuracy is 73.4%, greater than CMCT. However, in scene and object approach (Li and Fei-Fei, 2007), manual segmentation and object labels are used as additional inputs, a procedure that is not used in CMCT.

4.2.4 67-Class Indoor Scene Recognition

Following (Quattoni and Torralba, 2009), in this dataset, we use 80 images in each category for training and 20 images for testing. The experiments performed by (Quattoni and Torralba, 2009) with Gist achieved about 21% average recognition accuracy. When it is used local and global information to represent the scenes, the accuracy was improved to 25%. By using CMCT we achieve 25.82 ± 0.72%. The experiments performed using CENTRIST with no levels achieved 22.46 ± 0.84%. As one can see, in this challeging dataset, CMCT reaches better results than all techniques presented.

4.3 Experiments with GistCMCT

In all experiments performed in this section, we employed Histogram Intersection kernel (HIK) (Wu and Rehg, 2009) Support Vector Machine. For testing Gist we used the Matlab code provided by (Oliva and Torralba, 2001). The division of training and test is the same adopted in the previous experiments.

In the 15-Category Dataset, GistCMCT results is significantly greater than CMCT, since GistCMCT achieves 82.37 ± 0.11% accuracy. GistCMCT also outperforms Angular Radial Partitioning (ARP) Gist (Wei Liu and Gabbouj, 2012), a technique that improves Gist by modifying the grid division, which achieve 75.25 ± 0.67% accuracy on this dataset. By comparing confusion matrices presented in Figure 4 and Figure 5, the mean of diagonal matrix values increased from 0.77 to 0.82 when GistCMCT is used. Furthermore, it is possible to verify that the recognition rates in all outdoor classes are improved, as well as most recognition rates in indoor classes.

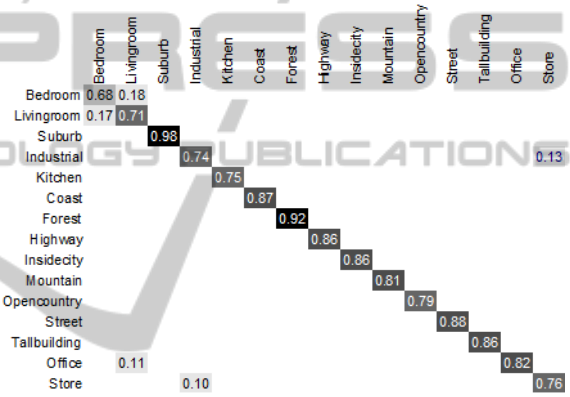

Figure 5: Confusion matrix from one run for 15-class scene recognition experiment using GistCMCT descriptor.

GistCMCT results is also significantly greater than CMCT in the 8-category, since GistCMCT achieves 85.64 ± 0.88% accuracy. ARP Gist (Wei Liu and Gabbouj, 2012), which achieves 84.77 ± 0.71% accuracy is, again, overcomed by GistCMCT in this dataset.

In the 8-class sports event GistCMCT achieves 76.08 ± 1.58% accuracy and, therefore, it overcomes CMCT and the results presented in (Li and Fei-Fei, 2007).

GistCMCT achieves 32.60 ± 0.81% in the outdoor scenes dataset showing better results than CMCT. It is interesting to note that even without outdoor scenes, adding Gist descriptor improves performance in this dataset.

5 CONCLUSIONS

The Contextual Mean Census Transform captures structural properties by modeling distribution of local

structures and combines it with contextual information. Those contextual information are obtained from the distribution of local structures formed from local structures in the original image, to perform scene recognition task. CMCT combines a modification of CENTRIST with contextual information. Comparing the results of MCT(8 bits) histogram and CMCT, one can see that the introduction of contextual information improves the image representation. Furthermore, CMCT preserves the advantages of CENTRIST (easy to implement, almost no parameter to tune, low illumination dependence) and shows better performance, as one can see in the presented experiments. As CENTRIST, CMCT is not invariant to rotation.

The GistCMCT, in its turn, is a combination of two holistic approach: Gist and CMCT. In Tables 1 and 2 it is possible to see that CMCT is overcomed by Gist when only outdoor scenes are presented and outperforms Gist when indoor scenes are classified. The combination of these two different global descriptors improves classification and outperforms as much Gist as CMCT. Besides the good performance, GistCMCT does not need creating codebooks, which is often computationally intense.

In our future research, we intend to use some form of associating spatial layout information, as subregions of different resolution levels, and include another type of information to improve the classification performance.

REFERENCES

Chang, C.-C. and Lin, C.-J. (2011). Libsvm: A library for support vector machines. *ACM Trans. Intell. Syst. Technol.*, 2(3):1–27.

Fei-Fei, L. and Perona, P. (2005). A bayesian hierarchical model for learning natural scene categories. In *Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05) - Volume 2*, pages 524–531. IEEE Computer Society.

Fröba, B. and Ernst, A. (2004). Face detection with the modified census transform. In *Proceedings of the Sixth IEEE international conference on Automatic face and gesture recognition*, pages 91–96. IEEE Computer Society.

Grauman, K. and Darrell, T. (2005). The pyramid match kernel: Discriminative classification with sets of image features. In *Proceedings of the Tenth IEEE International Conference on Computer Vision - Volume 2*, ICCV '05, pages 1458–1465. IEEE Computer Society.

Lazebnik, S., Schmid, C., and Ponce, J. (2006). Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Volume 2*, CVPR '06, pages 2169–2178. IEEE Computer Society.

Li, L.-J. and Fei-Fei, L. (2007). What, where and who? classifying events by scene and object recognition. In *IEEE 11th International Conference on Computer Vision*, pages 1–8. IEEE Computer Society.

Liu, S., Xu, D., and Feng, S. (2011). Region contextual visual words for scene categorization. *Expert Systems with Applications*, 38(9):11591–11597.

Lowe, D. G. (1999). Object recognition from local scale-invariant features. In *Proceedings of the International Conference on Computer Vision - Volume 2*, ICCV '99, pages 1150–1157. IEEE Computer Society.

Oliva, A. and Torralba, A. (2001). Modeling the shape of the scene: A holistic representation of the spatial envelope. *Int. J. Comput. Vision*, 42(3):145–175.

Oliva, A. and Torralba, A. (2006). Building the gist of a scene: The role of global image features in recognition. *Progress in Brain Research*, 155:23–36.

Qin, J. and Yung, N. (2010). Scene categorization via contextual visual words. *Pattern Recognition*, 43(5):1874–1888.

Quattoni, A. and Torralba, A. (2009). Recognizing indoor scenes. In *Proceedings IEEE CS Conf. Computer Vision and Pattern Recognition*, pages 413–420. IEEE Computer Society.

Salton, G. and McGill, M. J. (1983). *Introducrion to modern information retrieval*. New York: McGraw-Hill.

Vapnik, V. (1998). The support vector method of function estimation. *Nonlinear Modeling advanced blackbox techniques Suykens JAK Vandewalle J Eds*, pages 55–85.

Wei Liu, S. K. and Gabbouj, M. (2012). Robust scene classification by gist with angular radial partitioning. In *Communications Control and Signal Processing (ISCCSP), 2012 5th International Symposium on*, pages 2–4.

Wu, J. and Rehg, J. M. (2009). Beyond the euclidean distance : Creating effective visual codebooks using the histogram intersection kernel. In *Computer Vision, 2009 IEEE 12th International Conference on*, pages 630–637. IEEE Computer Society.

Wu, J. and Rehg, J. M. (2011). Centrist: A visual descriptor for scene categorization. *IEEE Trans. Pattern Anal. Mach. Intell.*, 33(8):1489–1501.

Zabih, R. and Woodfill, J. (1994). Non-parametric local transforms for computing visual correspondence. In *Proceedings of the third European conference on Computer Vision - Volume 2*, ECCV '94, pages 151–158. Springer-Verlag New York, Inc.