IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661, p- ISSN: 2278-8727Volume 14, Issue 4 (Sep. - Oct. 2013), PP 22-27
www.iosrjournals.org

Computer Vision: Visual Extent of an Object

Akshit Chopra[1], Ayushi Sharma[2]
[1](*Department Of Computer Science, Maharaja Surajmal Institute Of Technology – Guru Gobind Singh Indraprastha University, India)*
[2](*Department Of Computer Science, Maharaja Surajmal Institute Of Technology – Guru Gobind Singh Indraprastha University, India)*

Abstract: *The visual extent of an object reaches beyond the object itself. It is reflected in image retrieval techniques which combine statistics from the whole image in order to identify the image within. Nevertheless, it is still unclear to what degree and how this visual extent of an object affects the classification performance. Here we analyze the visual extent of an object on the Pascal VOC dataset using bag of words implementation with SIFT Descriptors. Our analysis is performed from two angles: (a) Not knowing the object location, we determine where in the image the support for object classification resides (normal situation) and (b) Assuming that the object location is known, we evaluate the relative potential of the object and its surround, and of the object border and object interior (ideal situation).*

Key words: *Computer vision, Content based image retrieval, Context, Extent of an Object, Visual extent*

I. Introduction:

COMPUTER VISION is a field that includes methods for acquiring, processing, analyzing, and understanding images and, in general, high-dimensional data from the real world in order to produce numerical or symbolic information, *e.g.*, in the forms of decisions. A theme in the development of this field has been to duplicate the abilities of human vision by electronically perceiving and understanding an image. This image understanding can be seen as the disentangling of symbolic information from image data using models constructed with the aid of geometry, physics, statistics, and learning theory. Computer vision has also been described as the enterprise of automating and integrating a wide range of processes and representations for vision perception.

In the early days of computer vision the visual extent of the object was sought to be precisely confined to its silhouette. And for good reasons as object boundaries (i) are more stable against lighting changes than the rest of the surface, (ii) indicate the object geometry directly, and (iii) reduce the processing requirements. This gave birth to the fact that an object should be accurately segmented before it can be recognized. However, the task of finding the contour- bounded location of an object is very hard to solve and not mainly necessary of object recognition. Recently, the use of powerful local descriptors, the ever increasing size of data sets to learn from and the advances in statistical pattern recognition have rendered obsolete the necessity to know the object location before object-based image classification. The first step on the road to less localization of the object was to use local region descriptors in a specific spatial arrangement. This allowed the object to be found based on only its discriminative features. The second step was the introduction of the Bag-of-Words method which selects interesting regions, converts them to visual words, and uses word counts followed by a spatial verification step to retrieve matching image regions. This is followed by the generalized Bag-of- Words to image classification and removed the spatial verification, relying on interest point detectors to extract visual words from the object. In the final step, the quantity of visual words was found to be more important than the quality of the location of the visual words.

II. Procedure

In Bag of Words a SIFT variant is used. Normally, SIFT divides a patch into 4 by 4 sub patches where for each sub-patch a Histogram of Oriented Gradients is calculated. Two lines of investigation are followed, as visualized in Fig. 1. The first line is the *normal* situation where we a visual concept detection algorithm is applied and which image parts contribute how much in identifying the target object are determined. The second line is the *ideal* situations where we use the known object locations to isolate the object, surround, and object interior and object border. The first line shows what currently *is* measured, and the second reveals what *could* be measured.

We evaluate the visual extent of an object in the Bag of-Words framework in terms of the object surround, object border, and object interior. First we investigate the influence of the surround with respect to the complete object. Then we investigate the influence of the object border with respect to the object interior. The ground truth object locations are used to isolate the object from its surround in both aspects. As the Bag-of-

Words framework thrives using lots of data, a large dataset where the locations are given in terms of bounding boxes is used. In the normal situation the distinction is made between object/surround and interior/border after classification on the test set only. In the ideal situation this distinction is made beforehand on both the training and test set. When there are multiple instances of the same class their measurements are combined to avoid measuring object features in its surround.

1.1 The Dataset

We choose to use datasets from the widely used Pascal VOC challenge as this allows for a good interpretation and comparison with respect to other work. We benchmark our Bag of-Words algorithm on the Pascal VOC 2007 classification challenge to show our framework is competitive. Our analysis is done on two Pascal VOC 2010 datasets. First, we use the classification dataset which provides the object locations in terms of bounding boxes. In this dataset we emphasize quantity of annotations over the quality of annotations. Second, we use the segmentation dataset which is much smaller but provides more accurate object locations in terms of segments. For the Pascal VOC 2010 datasets we use the predefined train set for training and the val set for testing.

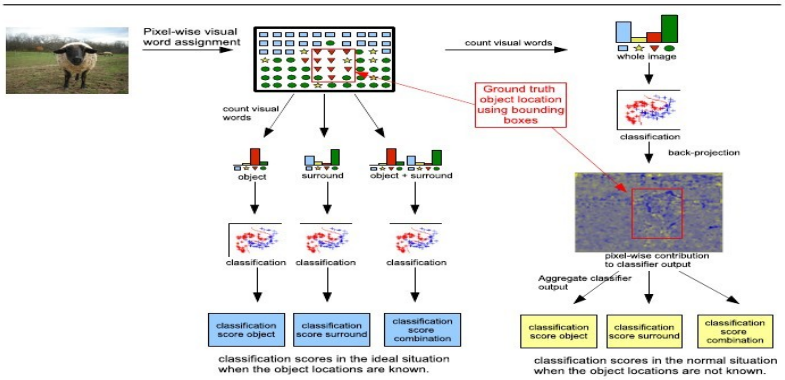

FIGURE 1

1.2 The Confusion Average Precision Matrix (CAMP)

To facilitate analysis, a confusion matrix is developed based on the Average Precision, which we call Confusion Average Precision Matrix or CAMP. The CAMP includes the Average Precision in its diagonal elements and, similar to a confusion matrix, shows which classes are confused. We define the confusion or off-diagonal elements of the CAMP as the total loss of Average Precision of encountering a specific non-target class in the ranked list. To calculate the loss we traverse the ranked list in decreasing order of importance. When a non-target class is encountered at position i, the loss L is the difference between the AP assuming a perfect ranking from position i and the AP assuming a perfect ranking from position $i + 1$.

The total confusion with a non-target class d is the sum of loss to that class, calculated by $\sum_{x_i \in d} L(x_i)$. As we measure

confusion in terms of loss, by definition the AP plus the sum of the loss over all classes adds to one.

1.3 Bag-of-Words Framework

A condense overview of our Bag-of-Words implementation (Uijlings et al. 2010) is given in Table 1. We sample small regions at each pixel which is an extreme form of sampling using a regular, dense grid (Jurie and Triggs 2005;

Nowak et al. 2006). From these regions we extract SIFT (Lowe 2004) and four colour SIFT variants (van de Sande et al. 2010) which have been shown to be superior for image retrieval (Mikolajczyk and Schmid 2005; van de Sande et al. 2010; Zhang et al. 2007). Thus we use intensity-based SIFT, opponent-SIFT, rg-SIFT (normalised RGB), RGBSIFT, and C-SIFT. Normally, SIFT consists of 4 by 4 subregions. However, we want our descriptors to be as small as possible in our experiments to be able to make the distinctions between object interior, object border, and object surround as crisp as possible.We therefore extract SIFT features of 2 by 2 subregions, which degrades performance no more than 0.02 MAP. The size of such SIFT patch is 8 by 8 pixels.

Table 1 Overview of our Bag-of-Words implementation. In our two lines of analysis we divide the image into subregions by either using the Spatial Pyramid or the ground truth object locations

Descriptor extraction	Word assignment	Classification
• Sampling each pixel • Size: 8×8 pixels • Descriptors: – 2×2 SIFT – 2×2 opp-SIFT – 2×2 rg-SIFT – 2×2 RGB-SIFT – 2×2 C-SIFT	• PCA dimension reduction by 50% • Random Forest: 4 binary decision trees of depth 10	• SVM: – Hist Int kernel • Image Divisions: Spatial Pyramid – 1×1, 1×3 Ground truth loc. – object/surround – interior/border

1.4 Analysis Without Knowing the Object Location

The line of analysis where the object locations are unknown shows how all parts of the image are used for classification by current state-of-the-art methods. We first classify images using a standard, state-of-the-art Bag-of-Words framework. After classification, we project the output of the classifier back onto the image to obtain a visualisation of pixel-wise classifier contributions; the sum of the pixel-wise contributions is equal to the output of the original classifier, which measures the distance to the decision boundary. After we have created the pixel-wise classifier contributions, we use the ground truth object locations to determine how much each image part (i.e. surround, object, object interior,object border) contributes to the classification. When an image contains multiple objects of the same class, we add contributions of all its locations together. When an image contains the target class, its location is used to make the distinction into object, surround, object interior, and object border. If the image does not contain the target class, we use the class with the highest classification contribution to make this distinction. This allows us to create a partitioning for both target and non-target images, which we need in order to calculate the Average Precision that is defined over the whole dataset (there is no "true" partitioning into the object and its surround for non-target images).

1.5 Distinguishing Object, Surround, Interior, and Border

For boxes, the ground truth locations separate the object from the surround. Note that the nature of the boxes cause some surround to be contained in the object. To separate the object interior from the object border, we define an object interior box as being a factor n smaller than the complete object box while its centre pixel remains the same. To determine the interior box we use the idea that object border contains the shape and the object interior contains texture and interior boundaries, which should be complementary. Separating complementary information should yield better results for the combination, hence we find the optimal interior boxes by optimising classification accuracy over n on the training set using cross-validation.We found a factor 0.7 to be optimal. This means that 49% of the object is interior and the rest border. For segments the Pascal VOC dataset only annotates the interior of the object while there is a 5 pixel zone around where the borders of the objects are. We want to ensure that no surround descriptors measure this border zone, and no interior descriptors measure this border zone. As we use the middle of our descriptor as point of reference in the ideal situation, we extend this border zone with half our descriptor size both inwards and outwards. Extending the border outwards yields our outlines of the object. Extending the border inwards yields the separation between object interior and object border. Our object border hence becomes 13 pixels wide. We measured that on average over all objects, 46% of the object becomes interior and the rest border.

III. Proposed Results

3.1 Classification Without Knowing the Object Location

We first benchmark our Bag-of-Words system on the Pascal VOC 2007 dataset, on which most results are published. For our normal Bag-of-Words system where we do not know the object location we achieve an accuracy of 0.57 MAP, sufficiently close to recent state-of-the-art Bagof- Word scores obtained by Harzallah et al. (2009) and van de Sande et al. (2010), which are respectively 0.60MAP and 0.61 MAP. To enable back-projection with (6) we use the Histogram Intersection kernel instead of the widely accepted $\chi 2$ kernel (Harzallah et al. 2009; Jiang et al. 2007; van de Sande et al. 2010; Zhang et al. 2007). This does not influence classification accuracy: with the $\chi 2$ kernel performance stays at 0.57 MAP. Instead, most of the difference in accuracy between our work and Harzallah et al. (2009), van de Sande et al. (2010) can be attributed to our use of 2×2 SIFT patches: using the four times as large 4×4 SIFT descriptor results in a classification accuracy of 0.59 MAP. However, in most of our experiments we favour small SIFT descriptors to minimise the overlap between object and surround, and interior and border descriptors. From now on all results are reported on the Pascal VOC 2010 dataset using 2×2 SIFT descriptors, unless otherwise noted.

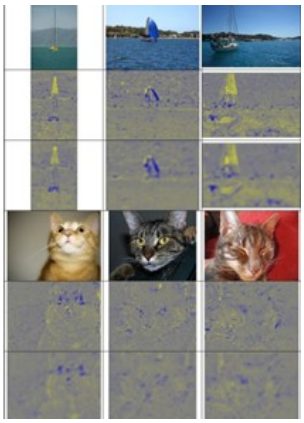

FIGURE 2 : Pixel-wise contribution to the classification for top ranked images for the categories *boat and cat*. The original image is followed by the contribution of 2×2 and 4×4 SIFT respectively. *Dark-blue* means a negative and *light-yellow* means a positive contribution to the classifier. Notice that high positive or high negative contributions are often located on small details. The 4 ×4 SIFT images resemble a blurred version of their 2×2 counterparts

3.2 Classification in Ideal Setting with Known Object Location

In this experiment we use the object location to create a separate representation for the surrounding and the object, where the representation of the object may be split into the interior and the border of the object. We compare this with the results of normal situation where the object location is not known. Clearly, for all classes knowledge of the object location greatly increases performance. The overall accuracy of the normal situation is 0.54 MAP, the accuracy of the ideal situation when making the distinction between object and surround is 0.68 MAP (where no Spatial Pyramid is applied to the object). When creating separate representations for the surround, object interior, and object border performance increases to 0.73 MAP. This shows that the potential gain of knowing the object locations is 0.19 MAP in this dataset. Similarly, on the segmentation dataset, in the normal situation where the object location is not known the classification accuracy is 0.44MAP.When separating the object from the surround accuracy rises to 0.62 MAP. If we make a separation between surround, object interior, and object border accuracy improves to 0.69 MAP. The huge difference between the accuracy without and without knowing the object location shows that the classifier cannot distinguish if visual words belong to the object or surround. We investigate the cause by determining for each visual word the probability that it occurs in an object (i.e. in any of the specified object classes). This graph shows that 1% of the words have a larger than 90% probability of describing background. We
found that these words describe mostly homogeneous texture (e.g. *sky*). In contrast, no single word has a larger than 90% probability of occurring on an object and less than 2% of the visual words occur on an object more than 75% of the cases. Note that these numbers are the same when using 4×4 SIFT. This means that no visual words exclusively describes objects and that these visual words are less specific than generally thought.

3.3 Object Versus Surround Using Segments

We repeated the experiments to analyse the influence of the object and the surround, but this time on fewer data but using more accurate object locations in terms of segments. The comparison of the influence between the object and its surround in the normal situation for segments looks similar, except that performance of using only the object is worse. Hence with fewer training examples the classifier is still able to learn the appearance of the surrounding but has less success in learning the appearance of the object itself. This means that the appearance of the context is simpler than that of an object. To verify whether this change in behaviour comes from the omission of any context while using segments, we repeated the experiment on the segmentation dataset but using boxes. Results were the same. Hence we conclude that the behaviour results from using fewer training examples: to accurately learn the appearance of these relatively difficult classes more training data is needed.

3.4 Interior versus Border

The object interior consists of texture and of interior boundaries, reasonably captured by a Bag-of-Words representation. However, this representation may be less appropriate for the object boundary as the object shape is intuitively better represented by larger fragments with more geometric constraints.

Hence while the conclusions made on the relative contribution of the border and the interior may not extend to object recognition in general, it will still be indicative of the relative difficulty of obtaining information of the object border and object interior.

IV. Conclusion:

This paper investigated the visual extent of an object in terms of the object and its surround, and in terms of the object interior and the object border. Our investigation was performed from two perspectives: The normal situation where the location of the objects are unknown, and an ideal situation with known object locations. For the normal perspective we visualised how the Bag-of-Words framework classifies images. These visualisations indicate that the support for the classifiers is found throughout the whole image occurring indiscriminately in both the object and its surround, supporting the notion that context facilitates image classification (Divvala et al. 2009; Oliva and Torralba 2007). While for some classes with a highly varying surround Bag-of-Words learns to ignore the context, as observed by Zhang et al. (2007), this does not generalise to all classes. We found that the role of the surroundings is significant for many classes. For *boat* the object area is even a negative indicator of its presence. At the same time, we have demonstrated that when the object locations are known a priori, the surroundings do not help to increase the classification performance significantly. After ideal localisation, regardless of the size of the object, the object appearance alone predicts its presence equally well as the combination of the object appearance and the surround.

We showed that no visual words uniquely describe only object or only surround. However, by making the distinction between object and surround explicit using the object locations, performance increases significantly by 0.20 MAP. This suggests that modelling the object location can lead to further improvements within the Bag-of-Words framework, where we see the work of Harzallah et al. (2009) as a promising start. Regarding the surround the following view arises. The surroundings are indispensable to distinguish between groups of classes: furniture, animals, and land-vehicles. When distinguishing among the classes within one group the surroundings are a source of confusion. Regarding the object features, we have observed differences how classes are being recognised: (1) For the physically rigid *aeroplane, bicycle, bus, car*, and *train* classes interior and exterior boundaries are important, while texture is not. (2) The non-rigid animals *dog, cat, cow,* and *sheep* are recognised primarily by their fur while their projected shape varies greatly. While SIFT feature values respond to interior boundaries, exterior boundaries, and texture at the same time, the recognition differences suggest that using more specialised features could be beneficial. Bag-of-Words with SIFT measure texture, interior object boundary fragments, and shape boundary fragments as local details. For identifying the context of an image this is adequate, especially considering that context partially consists of shapeless mass-goods such as grass, sky, or water. In contrast, for objects features more spatial consistency could help. This suggests that future features would render more improvements on recognising objects than on recognizing context. This is consistent with the observation by Biederman (1981) in human vision that objects viewed in isolation are recognised as easily as objects in proper context.

References:

[1]. Agarwal, S., Awan, A., & Roth, D. (2004). Learning to detect objects in images via a sparse, part-based representation. IEEE Transactions on Pattern Analysis and Machine Intelligence, 26(11), 1475–1490.
[2]. Bar, M. (2004). Visual objects in context. Nature Reviews. Neuroscience, 5, 617–629.
[3]. Biederman, I. (1981). On the semantics of a glance at a scene. In Perceptual organization (pp. 213–263). Hillsdale: Lawrence Erlbaum.
[4]. Bishop, C. M. (2006). Pattern recognition and machine intelligence. Berlin: Springer.
[5]. Blaschko, M. B., & Lampert, C. H. (2009). Object localization with global and local context kernels. In British machine vision conference.
[6]. Burl, M. C., Weber, M., & Perona, P. (1998). A probabilistic approach to object recognition using local photometry and global geometry. In European conference on computer vision.
[7]. Carbonetto, P., de Freitas, N., & Barnard, K. (2004). A statistical model for general contextual object recognition. In European conference on computer vision. Berlin: Springer.
[8]. Csurka, G., Dance, C. R., Fan, L.,Willamowski, J., & Bray, C. (2004). Visual categorization with bags of keypoints. In ECCV international workshop on statistical learning in computer vision, Prague.
[9]. Dalal, N., & Triggs, B. (2005). Histograms of oriented gradients for human detection. In IEEE conference on computer vision and pattern recognition.
[10]. Divvala, S. K., Hoiem, D., Hays, J. H., Efros, A. A., & Herbert, M. (2009). An empirical study of context in object detection. In IEEE conference on computer vision and pattern recognition.
[11]. Everingham, M., van Gool, L., Williams, C. K. I., Winn, J., & Zisserman, A. (2010). The pascal visual object classes (voc) challenge.
[12]. International Journal of Computer Vision, 88, 303–338.
[13]. Fergus, R., Perona, P., & Zisserman, A. (2003). Object class recognition by unsupervised scale-invariant learning. In IEEE conference on computer vision and pattern recognition.
[14]. Fulkerson, B., Vedaldi, A., & Soatto, S. (2009). Class segmentation and object localization with superpixel neighborhoods. In IEEE international conference on computer vision
[15]. Geurts, P., Ernst, D., & Wehenkel, L. (2006). Extremely randomized trees. Machine Learning, 63(1), 3–42.
[16]. Gould, S., Fulton, R., & Koller, D. (2009). Decomposing a scene into geometric and semantically consistent regions. In IEEE international conference on computer vision.
[17]. Harzallah, H., Jurie, F., & Schmid, C. (2009). Combining efficient object localization and image classification. In IEEE international conference on computer vision.
[18]. Hoiem, D., Efros, A. A., & Hebert, M. (2008). Putting objects in perspective.
[19]. International Journal of Computer Vision, 80, 3–15.

[20]. Jiang, Y. G., Ngo, C. W., & Yang, J. (2007). Towards optimal bag-offeatures for object categorization and semantic video retrieval. In
[21]. ACM international conference on image and video retrieval (pp. 494–501). New York: ACM Press.
[22]. Jurie, F., & Triggs, B. (2005). Creating efficient codebooks for visual recognition. In IEEE international conference on computer vision.
[23]. Lazebnik, S., Schmid, C., & Ponce, J. (2006). Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In IEEE conference on computer vision and pattern recognition, New York.
[24]. Lowe, D. G. (2004). Distinctive image features from scale-invariant keypoints. International Journal of Computer Vision, 60, 91–110.
[25]. Maji, S., Berg, A. C., & Malik, J. (2008). Classification using intersection kernel support vector machines is efficient. In IEEE conferenceon computer vision and pattern recognition.
[26]. Malisiewicz, T., & Efros, A. A. (2007). Improving spatial support for objects via multiple segmentations. In British machine vision conference, September 2007.
[27]. Malisiewicz, T., & Efros, A. A. (2009). Beyond categories: the visual memex model for reasoning about object relationships. In Neural information processing systems.
[28]. Marszałek, M., Schmid, C., Harzallah, H., & van de Weijer, J. (2007).
[29]. Learning representations for visual object class recognition. In ICCV Pascal VOC 2007 challenge workshop. Mikolajczyk, K., & Schmid, C. (2005). A performance evaluation of
[30]. local descriptors. IEEE Transactions on Pattern Analysis and Machine Intelligence, 27(10), 1615–1630.
[31]. Moosmann, F., Triggs, B., & Jurie, F. (2006). Fast discriminative visual codebooks using randomized clustering forests. In Neural information processing systems (pp. 985–992).
[32]. Nedovi´c, V., & Smeulders, A.W. M. (2010). Stages as models of scene geometry. IEEE Transactions on Pattern Analysis and Machine Intelligence, 32, 1673–1687.
[33]. Nowak, E., Jurie, F., & Triggs, B. (2006). Sampling strategies for bagof- features image classification. In European conference on computer vision.
[34]. Oliva, A., & Torralba, A. (2001). Modeling the shape of the scene: a holistic representation of the spatial envelope. International Journal of Computer Vision, 42(3), 145–175.
[35]. Oliva, A., & Torralba, A. (2007). The role of context in object recognition. Trends in Cognitive Sciences, 11, 520–527.
[36]. Rabinovich, A., Vedaldi, A., Galleguillos, C., Wiewiora, E., & Belongie, S. (2007). Objects in context. In International conference
[37]. on computer vision (pp. 1–8).
[38]. Shotton, J., Winn, J., Rother, C., & Criminisi, A. (2009). Textonboost for image understanding: multi-class object recognition and segmentation by jointly modeling texture, layout, and context. International Journal of Computer Vision, 81, 2–23.
[39]. Singhal, A., Luo, J., & Zhu, W. (2003). Probabilistic spatial context models for scene content understanding. In IEEE conference on
[40]. computer vision and pattern recognition.
[41]. Sivic, J., & Zisserman, A. (2003). Video Google: a text retrieval approach to object matching in videos. In IEEE international conference on computer vision.
[42]. Smeaton, A. F., Over, P. & Kraaij, W. (2006). Evaluation campaigns and TRECVID. In ACM SIGMM international workshop on multimedia information Retrieval.
[43]. Smeulders, A. W. M., Worring, M., Santini, S., Gupta, A., & Jain, R. (2000). Content-based image retrieval at the end of the early
[44]. years. IEEE Transactions on Pattern Analysis and Machine Intelligence, 22(12), 1349–1380.
[45]. Tahir, M. A., van de Sande, K., Uijlings, J., Yan, F., Li, X., Mikolajczyk, K., Kittler, J., Gevers, T., & Smeulders, A. (2008). UVA
[46]. and surrey@Pascal VOC 2008. In ECCV Pascal VOC 2008 challenge workshop.
[47]. Tuytelaars, T., & Schmid, C. (2007). Vector quantizing feature space with a regular lattice. In IEEE international conference on computer vision.Uijlings, J. R. R., Smeulders, A. W. M., & Scha, R. J. H. (2009). What is the spatial extent of an object? In IEEE conference on computer vision and pattern recognition.
[48]. Uijlings, J. R. R., Smeulders, A. W. M., & Scha, R. J. H. (2010, in press). Real-time visual concept classification. IEEE Transactions on Multimedia. http://dx.doi.org/10.1109/TMM.2010.2052027
[49]. Ullah, M. M., Parizi, S. N., & Laptev, I. (2010). Improving bag-offeatures action recognition with non-local cues. In British machine vision conference.
[50]. van de Sande, K. E. A., Gevers, T., & Snoek, C. G. M. (2010). Evaluating color descriptors for object and scene recognition. IEEE
[51]. Transactions on Pattern Analysis and Machine Intelligence, 32, 1582–1596.
[52]. Vedaldi, A., & Zisserman, A. (2010). Efficient additive kernels via explicit feature maps. In IEEE conference on computer vision and pattern recognition.
[53]. Wolf, L., & Bileschi, S. (2006). A critical view of context. International Journal of Computer Vision, 69, 251–261.
[54]. Zhang, J., Marszałek, M., Lazebnik, S., & Schmid, C. (2007). Local features and Kernels for classification of texture and object categories: a comprehensive study. International Journal of Computer Vision, 73(2), 213–238