Hybrid Fusion: Beyond Early and Late Fusion for Texture Classification

Fahad Shahbaz Khan, Joost van de Weijer, and Maria Vanrell

Computer Vision Centre/Computer Science Department Building O, Campus UAB, 08193 Bellaterra (Barcelona), Spain
E-mail:fahad@cvc.uab.es

Abstract

This paper presents a novel method for combining multiple attributes in order to classify the different categories. We start by providing a detail analysis of how to optimally fuse color and shape information for texture classification. For this reason we analyze the two existing approaches, called early and late fusion, and argue that both approaches are suboptimal for some classes. To overcome this shortcoming, we propose to merge the two approaches into a single combined early and late fusion representation of an image. We further propose to combine this new hybrid fusion with a texture representation in an efficient way. Experiments have been conducted on a large dataset of ten different image categories and the results show that all these three cues are important for the task of texture classification and our proposed method increases the overall performance significantly. *Keywords*: Color vocabulary, Texture Vocabulary, Texture Categorisation.

1 Introduction and Related Work

Images play a fundamental part in our daily communication and the large amount of pictures digitally available are not manageable by humans anymore. Visual categorization is a difficult task, interesting in its own right, due to large variations between images belonging to the same class. Many features such as color, texture, shape, and motion have been used to describe visual information for visual categorization. This paper focuses on the difficult problem of texture categorisation.

A still open research question within the bag-of-words context is how to optimally fuse different images cues, like color and shape, into a single bag-of-words representation. Initially many methods only used the shape feature, predominantly represented by SIFT [6],to represent the image [9], [1] and [5]. However, more recently the possibility of adding color information has been investigated [7], [12], [8]. Most of the recent works combine color and shape at an early stage focussing on the photometrically invariant properties of the the color descriptors. However, none of these methods provide a thorough analysis of the problem of what is the optimal approach to fuse shape and color.

Generally, the fusion of color and shape is carried out in the visual-vocabulary construction stage. Creating a visual vocabulary is a challenging task as the vocabulary should be able to describe widely varying classes. Some classes might have very distinctive color, some very characteristic texture patterns and some might be characterized by combining both features. There exist two main approaches to fuse color and shape into the bag-of-words representation. The first approach, called early fusion , involves fusing local descriptors together and creating one joint shape-color vocabulary. The second approach, called late fusion , concatenates histogram representation of both color and shape, obtained independently. Most of the existing methods use early fusion [7], [12], [8] . One of the few works which compares both early and late fusion for image classification is done by [10] where both early fusion and late fusion have been discussed.

To this end, the paper has been organized as follows. In section 2 Vocabularies for texture, shape and color are discussed. Afterwards, in section 3, fusing multiple vocabularies is discussed and hybrid fusion scheme along with a texture representation is proposed. Section 4 presents the experimental details like the classification algorithm, the dataset used and the classification settings. Detailed experiments are shown in section 5. Finally, we sum up the conclusions.

2 Vocabulary for Texture, Color and Shape

Visual features color, shape and texture are used to characterise visual keywords. In our approach LBP and SIFT are used to create a texture and shape vocabulary. Two options are considered to create a color vocabulary namely Hue and Color Naming values. The main essence of our work lies in the combination of these three features. In the next sections texture, shape and color vocabulries have been discussed in detail.

2.1 Texture Vocabulary

For human perception texture is an important visual category. Texture is one of the most common low level features and plays an important role for the character of region for digital images. There are many different ways of solving the problem of texture analysis. In this regard we investigate the use of LBP for creating a texture vocabulary since it is known to yield very good good performance in recent texture studies [4] and [2]. For our experiments we have investigated different variations of LBP while creating a texture vocabulary.

2.2 Shape Vocabulary

Shape is one of the most common low level features and local Shapes are often regarded as one of the most discriminent features shared by different instances of an image category. In object recognition the shape of an object plays a pivotal role in searching for similar image objects. There are many different ways of solving the problem of shape analysis. In this regard we investigate the use of SIFT for creating shape vocabulary since it is known to yield very good performance in recent studies [5], [3].

2.3 Color Vocabulary

A color vocabulary is created to represent the color aspects of an image. The measured color values vary significantly due to large amount of variations. In this work color histogram approach is used in the Hue, Saturation, Value (HSV) color space [12] and Color Naming values mentioned in the work of [11]. Given a set of cluster centers (visual words), each image is then represented by a K (The number of clusters, K, optimized for the dataset) dimensional normalized frequency histogram n (W/I). Where W denotes the visual words and I denotes the set of images. For clustering K-means method is used.

3 Fusing Multiple Vocabularies

After creating the color and shape vocabulary, both vocabularies are then combined in a flexible manner to achieve better performance. The discriminative power of each vocabulary varies for different classes. Some classes are distinguished by color and some by shape. We first anaylze both early and late fusion which is followed by our proposed approach of combining both early and late fusion. We further show that combining this hybrid fusion with a texture representation imporve the results significantly.

3.1 Early Fusion and Late Fusion

In early fusion, the local features of color and shape are combined before quantization. The fusion involves concatenating the color features and shape features. From these combined vectors a joint vocabulary is obtained using the K-means algorithm. A weight vector β is introduced to tune the relative weight of the color and texture in the combined vocabulary V_{sc} .

$$V_{sc} = (\beta \ V_c, (1-\beta)V_s) \qquad (1)$$

where V_c are the color features and V_s are the texture features. The weight vector β is learned through cross-validation on the training data.

In late fusion the two features color and shape are computed independently. The two features are then fused together in one representation. Here the different vocabularies are concatenated after quantization. A weight vector α is introduced to obtain a combined histogram $n(w|I)$ of color and shape vocabularies for an image I.

$$n(w_{s\&c}|I) = \begin{bmatrix} \alpha \ n(w_c|I) \\ (1-\alpha) \ n(w_s|I) \end{bmatrix} \qquad (2)$$

where w is the number of total vocabulary words, w_c are Color words and w_s are shape words. The weight vector α is learned through cross-validation on training data.

Figure 1 highlights the two approaches used for combining the color and shape vocabulary.

3.2 Hybrid Fusion: Combining Early and Late Fusion

The work of [10] compares early versus late fusion. Both methods have their advantages and disadvantages. Early fusion obtains a vocabulary with a higher discriminative power, since the visual words describe both color and shape. Early fusion visual words could include red blobs, green lines, etc. . In late fusion the vocabularies are obtained by separately clustering the two cues shape and color. The image is thus represented as a distribution over shape-words and color-words. For example, if the image contains blobs and lines, and red and green features then from the late fusion representation it cannot be inferred that the image contains red lines or green blobs. In case of a class which is constant over both shape and color, early fusion representation is better. However for classes where only one of the cues is constant, late

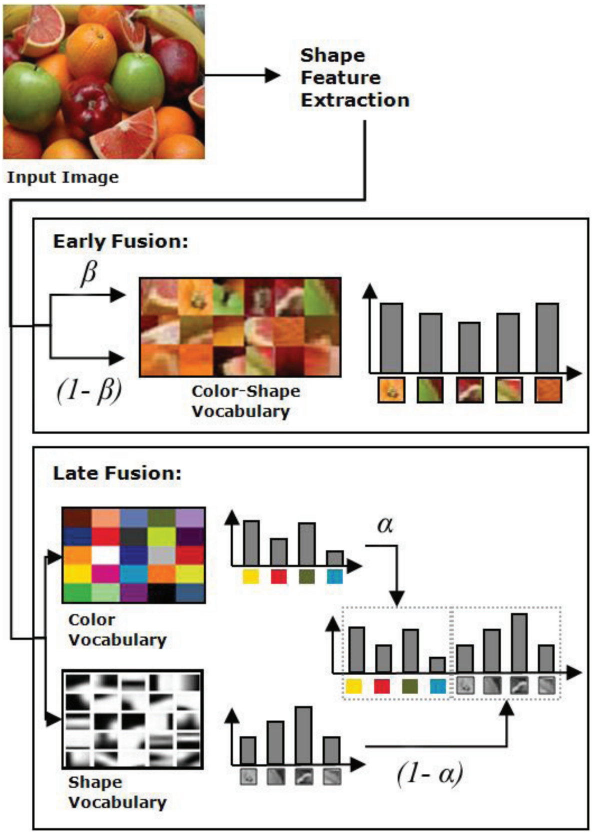

Figure 1: A Graphical explanation of early and late fusion schemes to combine color and shape information. The α and β parameters determine the relative weight of the two cues.

fusion representation is preferred. To combine the advantages of the two representations we propose to combine early and shape fusion into a single histogram. This will be done in a late fusion manner as described in Eq. 2.

3.3 Combining Hybrid Fusion with Texture Vocabulary

We take one step further in fusing multiple features by combining the hybrid fusion with a texture vocabulary. In our experiments we have used LBP for the creation of texture vocabulary. As a next step, different variations of LBP has been tested since the primitive LBP representation proved unsuccessful for our data set. Thus the final histogram is a weighted combination of late and early fusion of color and shape with texture histogram.

4 Experimental Setup

The performance of combined vocabularies will be tested on a the classification task using SVM. Details of the proposed procedure are outlined in this section.

4.1 Dataset

The approach outlined above is tested on a dataset with 10 classes (Marble, Wood, Beads, Foliage, Graffiti, Lace, Clouds, Fruit, and Water) with 40 images for each class. The images in the dataset have been collected from Google, Flickr, and Corel image collection. Figure 8 shows some of the images from the dataset. The dataset is very challenging due to wide range of textures and color in it. For example, the Foliage class in our dataset has mostly green color, but there are few images in this class that has red color in it. Similary there is a wide range of different texture patterns and colors in Marble and Graffiti class. There are a lot of variations in scale in the lace category.

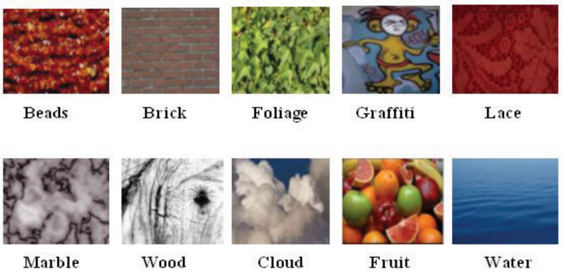

Figure 2: Typical examples of each class from the data set.

4.2 Classification Settings

The dataset has been divided into train set, validation set and test set. 25 images from each class are used for training and remaining 15 are used for testing.We believe the relative performance differences between various approaches to fuse multiple features to be independent of the detection method used. Thus an often used grid detector is employed where the patch centers lie 5 pixels apart. In our experiments multiclass non linear SVM with χ^2 kernel is used since it is known to produce best classification results [1]. To evaluate the classification performance we use the classification score. The classification score gives the percentage of correctly classified instances in the testset.

5 Experiments

This section explains in detail the creation of multiple vocabularies and the proposed methodology used for combining these vocabularies. Experiment 1 is about optmizing the individual vocabularies of texture. Experiment 2 provides an insight of color and shape vocabularies. Experiments 3 deals with combining both vocabularies in early and late fusion manner in order to optimize the classification performance. Finally in experiment 4 we combine the late and early fusion together in one representation. We further combine the hybrid and texture in one representation.

5.1 Experiment 1: Texture Vocabularies

This section provides detailed results obatined using only the texture information. As a first step a texture vocabulary has been created using local binary patterns. We explored different variations related to LBP such as Rotation Invariant LBP $\left(LBP^{ri}_{P,R}\right)$, Uniform LBP $\left(LBP^{u2}_{P,R}\right)$, Rotation Invariant with Uniform LBP $\left(LBP^{riu2}_{P,R}\right)$, Rotation Invariant Variance LBP $(VAR_{P,R})$, and Joint distribution of Rotation Invariant Uniform LBP with its Variance $\left(LBP^{riu2}_{P,R}/VAR_{P,R}\right)$.

LBP operator	P,R	Bins	Accuracy
$LBP^{ri}_{P,R}$	$8,1$	36	54.16
$LBP^{u2}_{P,R}$	$16,2$	243	62.30
$LBP^{riu2}_{P,R}$	$8,1$	10	47.27
$LBP^{riu2}_{P,R}/VAR_{P,R}$	$16,2/8,1$	328	58.10
$LBP^{ri}_{P,R}+LBP^{u2}_{P,R}$	$8,1+16,2$	279	64.27

Table 1: Classification Score (percentage) using LBP.

5.2 Experiment 2: Color and Shape Vocabularies

In this experiment we evaluated individual color and shape vocabularies. The results shows that for the categories in this dataset shape is a more important cue than color.

Vocabulary	Vocabulary Size	Accuracy
$SIFT$	700	73
HUE	400	56
$ColorNaming$	400	58

Table 2: Classification Score (percentage) using Shape and Color Vocabularies.

5.3 Experiment 3: Early Fusion and Late Fusion

In this experiment we combined shape and color vocabularies. In Table 3 the results of these experiments are summerised. The results using late fusion show a better classification score as compared to early fusion.

Vocabulary	Voc Size	Accuracy
$EarlyFusion(SIFT, HUE)$	1200	75
$LateFusion(SIFT, HUE)$	1100	77
$EarlyFusion(SIFT, CN)$	1200	77
$LateFusion(SIFT, CN)$	1100	79

Table 3: Classification Score (percentage) of Early and Late Fusion. Note that in both color cues (HUE and Color Names) late fusion performs better than early fusion.

5.4 Experiment 4: Combining Late and Early Fusion with Texture Vocabulary

The texture and shape/color are now combined by concatenating the histogram representation (late and early fusion) with LBP representation. The hybrid fusion of shape/color(color names) is denoted by *Hybrid(CN)* and shape/color(Hue) by *Hybrid(HUE)*.

Vocabulary	Vocabulary Size	Accuracy
$Hybrid(CN)$	2300	81
$Hybrid(HUE)$	2300	79
$Hybrid(CN)andLBP$	2310	83
$Hybrid(HUE)andLBP$	2310	81

Table 4: Classification Score (percentage) of hybrid fusion and combining hybrid fusion with LBP. Note that hybrid combination of Color Names with LBP provides the best results.

6 Discussion and Conclusions

The work presented in this paper is about classification on a large dataset of ten different texture categories using texture, shape and color featues. We investigated the two popular approaches of combining color and shape namely early and late fusion. Our results show that both late and early fusion have some short commings. The success of late fusion over early fusion and vice-versa depends on the nature of the categories in the dataset. In our case late fusion perform slightly better than early fusion. This could be due to the fact that in most of the categories only one of the cues, either color or shape, is constant. For example, foliage class is difficult to classify based on shape but relatively stable with respect to color appearance. In this case it is hard to find visual words based on early fusion that are consistent. As a next step late fusion has been analysed deeply by trying different combinations of vocabularies.

All three cues, color, shape and texture are found to be crucial to obtain a good overall classification score. Texture alone obtained 64.27%, color alone 58%,and shape alone provided 73% scores. The final combination got 83% which is obtained by combining hybrid fusion with LBP. When combining the vocabularies, color improves texture classification performance but best performance is achieved when shape has more influence than color.

Finally our final results suggest that all three cues are important for texture classification. Moreover fusing color and shape one step beyond early and late fusion improved the results. Combining our proposed hybrid fusion scheme with LBP respresentation further improved the overall classification score. The results also goes to show that combining multiple vocabularies clearly outperforms the performance of individual vocabulary cues.

References

[1] J. Zhang, M. Marszalek, S. Lazebnik, and C. Schmid, ”Local features and kernels for classification of texture and object catergories: A Comprehensive Study”, *International Journal of Computer Vision*, 73(2): 213-238, 2007.

[2] Topi Maenpaa, Matti Pietikainen, ”Classification with color and texture: jointly or separately?”, *Pattern Recognition*, 37(8): 1629-1640, 2004.

[3] K. Mikolajczyk and C. Schmid, ”A performance evaluation of local descriptors”, *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 27(10):16151630, 2005.

[4] T. Ojala, M. Pietikinen, and T. Menp, ”Multiresolution gray-scale and rotation invariant texture classification with local binary patterns”, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24:971-987, 2002.

[5] S. Lazebnik, C. Schmid, and J. Ponce, ”A Sparse Texture Representation Using Local Affine Regions”, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2005.

[6] D. G. Lowe, ”Distinctive image features from scale-invariant points”, *International Journal of Computer Vision*, 60(2):91-110, 2004.

[7] A. Bosch, A. Zisserman, and J.Munoz, ”Scene classification via plsa”, *In Proc. ECCV*, 2006.

[8] K. van de Sande, Th. Gevers, and C. Snoek, ”Evaluation of color descriptors for object and scene recognition”, *In Proc. CVPR*, 2008.

[9] G. Dorko, C. Schmid, ”Selection of scale-invariant parts for object class recognition”, *In Proc ICCV*, 2003.

[10] P. Quelhas and Jean-Marc Odobez, ”Natural scene image modeling using color and texture visterms”, *In Proc. CIVR*, 2006.

[11] J. van de Weijer, C. Schmid, and J.J. Verbeek, ”Learning color names from real-world images”, *In Proc. CVPR*, Minneapolis, Minnesota, USA, 2007.

[12] J. van de Weijer, C. Schmid, ”Coloring local feature extraction”, *In Proc. of the European Conference on Computer Vision*, volume 2, pages 334-348, Graz, Austria, 2006.