Grouplet: A Structured Image Representation for Recognizing Human and Object Interactions

Bangpeng Yao Li Fei-Fei
Computer Science Department, Stanford University, USA
{bangpeng,feifeili}@cs.stanford.edu

Abstract

*Psychologists have proposed that many human-object interaction activities form unique classes of scenes. Recognizing these scenes is important for many social functions. To enable a computer to do this is however a challenging task. Take people-playing-musical-instrument (PPMI) as an example; to distinguish a person playing violin from a person just holding a violin requires subtle distinction of characteristic image features and feature arrangements that differentiate these two scenes. Most of the existing image representation methods are either too coarse (e.g. BoW) or too sparse (e.g. constellation models) for performing this task. In this paper, we propose a new image feature representation called "*grouplet*". The grouplet captures the structured information of an image by encoding a number of discriminative visual features and their spatial configurations. Using a dataset of 7 different PPMI activities, we show that grouplets are more effective in classifying and detecting human-object interactions than other state-of-the-art methods. In particular, our method can make a robust distinction between humans playing the instruments and humans co-occurring with the instruments without playing.*

1. Introduction

In recent years, the computer vision field has made great progress in recognizing isolated objects, such as faces and cars. But a large proportion of our visual experience involves recognizing the interaction between objects. For example, seeing a human playing violin delivers a very different story than seeing a person chopping up a violin - one is a musician, the other is probably a contemporary artist. Psychologists have found that different brain areas are involved in recognizing different scenes of multiple objects [17] and in particular, there are neurons that react strongly upon seeing humans interacting with objects [15]. Such evidence shows that the ability to recognize scenes of human-object interactions is fundamental to human cognition.

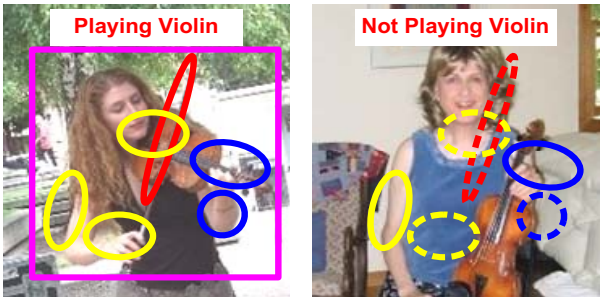

Figure 1. Recognizing a person playing violin versus not playing violin requires subtle discriminations of image features. Our algorithm discovers discriminative features called *grouplets* that encode rich, structured information for such tasks. In the **left** figure, three sample grouplets are shown in three different colors. Note that one grouplet (e.g. the cyan one) is represented by multiple image patches and their spatial configurations. In the **right** figure, we show that some information of the grouplets from the left is missing (their hypothetical locations are indicated by dashed lines), prompting our algorithm to decide that the person is not playing violin.

The goal of our work is to use structured visual features to recognize scenes in which a person is interacting with a specific object in a specific manner, such as playing musical instruments. Humans can recognize such activities based on only static images, most likely due to the rich structured information in the activities. For example, "playing violin" is defined not only by the appearance of a human and a violin and their co-occurrence, but also by the gesture of arms interacting with the pose of the violin, as shown in Fig.1.

One intuitive approach for this problem is to design an algorithm that can recognize the human pose, the target object, and the spatial relationship between the human and the object [27]. It is, however, an exceedingly difficult task to recognize complex human gestures. Most of the human pose estimation algorithms today cannot reliably parse out body parts that are crucial to our task, especially with partial occlusions or in cluttered backgrounds [23]. The same is also true for object detection. Detection rates of generic objects in realistic scenes are still low [5].

In this paper, instead of exploring models for pose estimation or object detection, we approach the problem by discovering image features that can characterize well differ-

ent human-object interactions. We take the view in [17] that such human-object configurations are like different types of scenes. So similar to scene and object classification [9, 18], our features need to discover different classes of activities that carry intrinsically different visual appearance and spatial information. This problem offers us an opportunity to explore the following issues that have not been widely studied in generic object recognition tasks:

- Spatial relations among image patches. Recognizing that a person is playing violin is not simply discovering the co-occurrence of the violin and the human, which could also occur when a person just standing next to a violin. Our features need to capture the spatial relations that are crucial to define the human-object interactions.
- More subtle and discriminative features. Most of the current image features (and models) are tested on classes of objects that are very different from each other (e.g. bicycles vs. cows). The classes of human-object interactions are much more similar, due to the dominant presence of humans in all classes. This demands more discriminative features to encode the image differences.

Focusing on the above issues, we propose a new image representation that encodes appearance, shape, and spatial relations of multiple image patches, termed "grouplet". The grouplets are discovered through a novel data mining approach, and could be further refined by a parameter estimation procedure. We show that the methods using grouplets outperform the state-of-the-art approaches in both human-object interaction *classification* and *detection* tasks.

The rest of this paper first presents a human-object interaction data set in Sec.2. Sec.3 and Sec.4 define the grouplets and introduce a method of obtaining discriminative grouplets respectively. Sec.5 briefly describes the classification methods that use grouplets. Related work is discussed in Sec.6. Experiment results are reported in Sec.7.

2. The PPMI Dataset

Most of the popular image data sets are collected for recognizing generic objects [6, 5] or natural scenes [21] instead of human and object interactions. We therefore collected a new data set called People-playing-musical-instruments (PPMI, Fig.2). PPMI[1] consists of 7 different musical instruments: bassoon, erhu, flute, French horn, guitar, saxophone, and violin. Each class includes ∼150 PPMI+ images (humans playing instruments) and ∼150 PPMI- images (humans holding the instruments without playing). As Fig.2 shows, images in PPMI are highly diverse and cluttered.

We focus on two problems on this data. One is to classify different activities of humans playing instruments; the other is to distinguish PPMI+ and PPMI- images for each instrument. The latter task is very different from traditional image classification tasks. Distinguishing PPMI+ and PPMI- images of the same instrument strongly depends on the structural information in the images, such as the spatial relations between the object and the human. This property of our data set cannot be captured by [12] and [13], which are possibly the only existing data sets of human-object interactions. Besides classification, we also show results of detecting people playing different instruments on the PPMI dataset.

[1] Resources of the images include image search engines Google, Yahoo, Baidu, and Bing, and photo hosting websites Flickr and Picassa.

Figure 2. Example images of the PPMI dataset. For each instrument, we show 3 images of people playing instruments (PPMI+) and 3 images of people co-occurring with but not playing the instruments (PPMI-).

3. Image Building Block - the Grouplet

For recognizing human-object interactions, we discover a set of discriminative features that encode the structured image information. To address the two central issues introduced in Sec.1, the grouplets have the following properties.

- Each grouplet contains a set of highly related image patches. It encodes the appearance, location, and shape of these patches, as well as their spatial relationship.
- For differentiating human and object interactions, we apply a novel data mining approach to discover a large number of discriminative grouplets.

A grouplet is defined by an AND/OR [4] structure on a set of *feature units*. A feature unit, denoted by $\{A, x, \boldsymbol{\sigma}\}$, indicates that a codeword of visual appearance A is observed in the neighborhood of location x (relative to a reference point). The *spatial extent* of A in the neighborhood

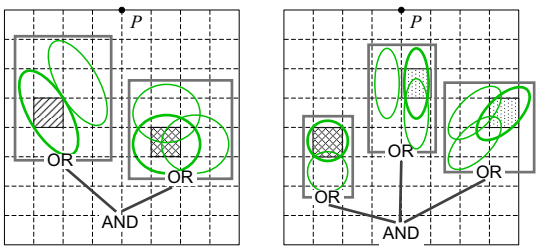

Figure 3. Two examples of grouplets: **left** is a size 2 grouplet; and **right** is a size 3 grouplet. Each grouplet lives in an image space where P indicates a reference location. Each grouplet is composed of a set of feature units. A feature unit, whose visual appearance is denoted by a shaded square patch, can shift around in a local neighborhood (indicated by smaller rectangular boxes). An ellipse surrounding the center of a feature unit indicates the spatial extent of the feature. Within the neighborhood, an OR operation is applied to select the feature unit that has the strongest signal (v, see Sec.4.1), indicated by the ellipse of thicker lines. An AND operation collects all feature units to form the grouplet.

of x is expressed as a 2D Gaussian distribution $\mathcal{N}(x, \boldsymbol{\sigma})$.

Fig.3 illustrates two grouplet features. Each ellipse denotes one feature unit. A grouplet is formed by applying some OR operations and an AND operation to a set of feature units. Each OR operation is applied to the feature units that have similar visual appearance and spatial extents, from which the one that has the strongest signal in the image is selected (thicker ellipses in Fig.3). The AND operation is applied to these selected feature units. The *size of a grouplet* is the number of OR operations it contains.

In the grouplet representation, each feature unit captures a specific appearance, location, and spatial extent information of an image patch. Together, the AND operation allows the grouplets to represent various interactions among a set of image patches, and the OR operation makes the grouplets resistent to small spatial variations. By definition, we do not exert any constraint on the appearance or location of the feature units, nor the size of the grouplets. Furthermore, the spatial extent of each feature unit will be automatically refined through a parameter estimation step (Sec.4.2.2), thus the grouplets can reflect any structured information among any number of image patches with any appearance. Examples of grouplets are shown in Fig.1 and Fig.10.

Implementation Details: In the grouplet representation, SIFT descriptors [19] are computed over a dense image grid of D rectangular patches, as in [18]. Using k-means clustering, we obtain a SIFT codebook which contains 250 codewords. Therefore, the visual appearance can be represented by $\{A_w\}_{w=1}^{W}$, where W=250. The feature units in one OR operation should have the same visual codeword. Reference points are chosen as the centers of the human faces.

4. Obtaining Discriminative Grouplets

To recognize subtly different scenes, we would like to find a rich set of grouplets that are not only highly characteristic of the image class, but also highly discriminative compared to other classes. We propose a novel data mining algorithm for discovering discriminative grouplets.

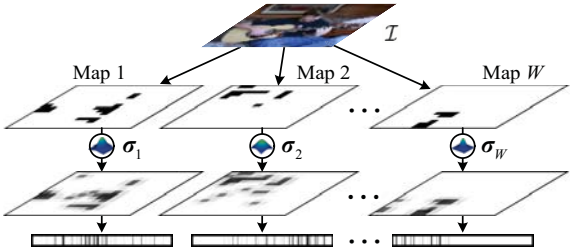

Figure 4. Computing the signals v of all feature units on an image $\mathcal{I}$. First, a codeword assignment map is obtained for each codeword A_w. In Map w, a region is marked black if it is assigned to A_w. Then, each Map w is convolved with a 2D Gaussian distribution with covariance $\boldsymbol{\sigma}_w$. Finally the results are concatenated into a $(D\times W)$-dimensional vector of signal values, where each entry is the signal value of a feature unit on the image.

4.1. Defining Discriminative Grouplets

Grouplet Λ is discriminative for class c means that Λ *has strong signals on images of class c, and has weak signals on images of other classes*. In the rest of this section, we first describe how to compute the signal values of feature units and grouplets, and then elaborate on the definition of discriminative grouplets.

The signal v of a feature unit $\{A, x, \boldsymbol{\sigma}\}$ on an image $\mathcal{I}$ is the likelihood that $\{A, x, \boldsymbol{\sigma}\}$ is observed in $\mathcal{I}$:

$$v = \sum_{x' \in \Omega(x)} \Big[p(A|a') \cdot \mathcal{N}(x'|x, \boldsymbol{\sigma}) \Big] \qquad (1)$$

where $\Omega(x)$ is the image neighborhood of location x, a' is the appearance of the image patch at x', $p(A|a')$ is the probability that a' is assigned to codeword A. Please refer to Fig.4 and implementation details of this section for more details. For a codeword A_w, we use a single variance $\boldsymbol{\sigma}_w$ to encode its spatial distribution in all positions of the image.

Given the signal values of the feature units in a grouplet, each OR operation selects a feature unit that has the strongest signal (see Fig.3). The overall signal of the grouplet, i.e. result of the AND operation, is the smallest signal value of the selected feature units. Intuitively, this decision ensures that even the relatively weakest feature unit needs to be strong enough for the grouplet to be strong (see Fig.5). In order to evaluate the discriminability of a grouplet, we introduce two terms, *support value*, $Supp(\cdot)$ and *confidence value*, $Conf(\cdot)$. A grouplet Λ is discriminative for a class c if both $Supp(\Lambda, c)$ and $Conf(\Lambda, c)$ are large. Given a set of training images where the signal of Λ on image $\mathcal{I}_i$ is denoted as r_i, $Supp(\Lambda, c)$ and $Conf(\Lambda, c)$ are computed by

$$Supp(\Lambda, c) = \frac{\sum_{c_i=c} r_i}{\sum_{c_i=c} 1}, \; Conf(\Lambda, c) = \frac{Supp(\Lambda, c)}{\max_{c' \neq c} Supp(\Lambda, c')} \qquad (2)$$

Figure 5. Example grouplets whose feature units are of different signal value strengths. One grouplet is presented in each image, where the ellipses indicate the location and spatial extent of the feature units. Thicker lines indicate stronger signal values. For the same flute-playing activity, it is intuitive to see that the grouplet on the **left** has overall stronger feature units than the mixed one in the **middle** and the weaker one on the **right**.

where c_i is the class label of $\mathcal{I}_i$. Intuitively, a large $Supp(\Lambda, c)$ indicates that Λ generally has strong signals on images of class c, and a large $Conf(\Lambda, c)$ implies relatively weak signals of Λ on images of classes other than c.

Implementation Details: The size of $\Omega(x)$ is 5×5 patches. We assign each image patch to its nearest codeword: $p(A|a)$=1 if and only if A is a's nearest codeword. We initialize $\boldsymbol{\sigma}_w$ to $[0.6, 0; 0, 0.6]$ for any A_w. $\boldsymbol{\sigma}_w$ will be updated in the parameter estimation step (Sec.4.2.2).

4.2. A Novel Iterative Mining Algorithm

For each class, our goal is to find all the grouplets of large support and confidence values. One way is to evaluate these values on all possible grouplets. Assuming an image of D patches and a codeword vocabulary of size W, there are $D\times W$ possible feature units. The total number of grouplets is therefore $O(2^{D\times W})$ (in this paper $D\times W$=240250). Clearly, evaluating $Supp(\cdot)$ and $Conf(\cdot)$ of all the grouplets for each class is computationally infeasible.

We therefore develop a data mining algorithm for this task, which discriminatively explores the AND/OR structure of the grouplets in an Apriori mining [1] process. Furthermore, we introduce a novel parameter estimation method to better estimate the spatial distribution $\boldsymbol{\sigma}_w$ of each codeword A_w as well as to obtain a set of weights for the grouplets of each class. Our mining algorithm then iterates between the mining process and the parameter estimation process. An overview of the algorithm is shown in Algorithm 1, where l-grouplets indicate the grouplets of size l. We briefly describe the mining and the parameter estimation method in the rest of this section.

4.2.1. The Modified Apriori Mining Algorithm In each iteration of Algorithm 1, given the spatial distribution $\boldsymbol{\sigma}_w$ of each codeword A_w, we compute the signal values of all the feature units on each image as in Fig.4. We are then ready to mine the discriminative grouplets for every class.

We modify the Apriori [1] mining method to explore the AND/OR structures to select the discriminative grouplets. The main idea of Apriori mining is compatible with the AND operation: if an l-grouplet has a large support value, then by removing the feature units in any of its OR operations, the remaining $(l-1)$-grouplets also have large support values. Therefore, we can generate l-grouplets based only on the mined $(l-1)$-grouplets, instead of considering all the possibilities. The OR operation is used to obtain the 1-grouplets. For each codeword, a hierarchical clustering is applied to the feature units that have large enough support values. Each cluster is then initialized as a 1-grouplet. The mining process is briefly shown in Algorithm 1.

foreach *Iteration* **do**
- • Compute signals of all feature units on each image;
- **foreach** *Class* **do**
  - ⋆ Obtain the feature units whose $Supp(\cdot) > T_{Supp}$;
  - ⋆ Generate 1-grouplets; Set $l = 2$;
  - **while** *The number of* $(l-1)$*-grouplets* $\geqslant 2$ **do**
    - Generate candidate l-grouplets; Remove l-grouplets whose $Supp(\cdot) < T_{Supp}$; $l = l + 1$;
  - **end**
  - ⋆ Remove the grouplets whose $Conf(\cdot) < T_{Conf}$.
- **end**
- • Parameter estimation to refine $\boldsymbol{\sigma}_w$ for each A_w and obtain a weight for each mined grouplet.

end

Algorithm 1: Obtaining discriminative grouplets.

Implementation Details: The hierarchical clustering is based on the maximum distance metric, of which the threshold is two times the patch size. The mining algorithm automatically adjusts the values of T_{Supp} and T_{Conf} for each class, so that the number of mined grouplets for different classes are approximately the same. More details of the mining method can be found in Sec.A of this paper.

4.2.2. Refining Grouplets Given a set of mined grouplets, we introduce a parameter estimation method to further refine the spatial distribution $\boldsymbol{\sigma}_w$ of each codeword A_w. With the refined $\boldsymbol{\sigma}$, one can expect that more accurate signal values of the feature units can be computed, which in turn can be put into the mining process to obtain better grouplets in the next iteration. Furthermore, the algorithm computes a weight on each mined grouplet for each class. The combination of grouplets and the class-dependent weights can then be directly used for classification tasks (see Sec.5).

Given an image $\mathcal{I}$ with class label c, we compute the likelihood of $\mathcal{I}$ given a set of parameters $\boldsymbol{\theta}$, where $\boldsymbol{\theta}$ contains the parameters for the spatial extent of each codeword and the importance of each grouplet.

$$p(\mathcal{I}, c|\boldsymbol{\theta}) = p(c|\boldsymbol{\theta}) \sum_m \left[p(\mathcal{I}|\Lambda^m, \boldsymbol{\theta}) p(\Lambda^m|c, \boldsymbol{\theta})\right] \qquad (3)$$

where Λ^m indicates the m-th mined grouplet. $p(\mathcal{I}|\Lambda^m, \boldsymbol{\theta})$ denotes the likelihood of $\mathcal{I}$ given Λ^m. $p(\Lambda^m|c, \boldsymbol{\theta})$ models the importance of Λ^m for class c. We use an expectation-maximization (EM) algorithm to estimate the parameters $\boldsymbol{\theta}$. Due to space limitation, we elaborate the details of

the parameter estimation method in Sec.B. On a PC with a 2.66GHz CPU, our algorithm can process around 20000 grouplets under 3 minutes per EM iteration.

5. Using Grouplets for Classification

Having obtained the discriminative grouplets, we are ready to use them for classification tasks. In this paper, we show that grouplets can be used for classification either by a generative or a discriminative classifier.

A Generative Classifier. Recall that in Sec.4.2.2, our probabilistic parameter estimation process outputs the importance of each grouplet for each class. This can, therefore, be directly used for classification. Given a new image $\mathcal{I}$, its class label c is predicted as follows,

$$c = \arg\max_{c'} p(c'|\mathcal{I}, \boldsymbol{\theta}) = \arg\max_{c'} p(c', \mathcal{I}|\boldsymbol{\theta}) \qquad (4)$$

A Discriminative Classifier. Discriminative classifiers such as SVM can be applied by using groupets. Given an image, the input feature vector to SVM classifiers is the signal values of the mined grouplets.

6. Related Work

Many features have been proposed for various vision tasks in the past decade [26]. It is out of the scope of this paper to discuss all of them. Instead, we discuss the image representations that have directly influenced our work.

One of the most popular image feature representation schemes is bag of words (BoW) and its derivations (e.g. [18]). These methods have shown promising results in holistic image classification tasks. But by assuming little or no spatial relationships among image patches, these representations are not sufficient for more demanding tasks such as differentiating human and object interactions.

In order to remedy BoW, some methods have been proposed to either encode longer range image statistics [25, 24] or explicitly model spatial relationships among image patches [9, 7, 20]. But most of such approaches uncover image features in a generative way, which might result in some features that are not essential for recognition. In [8], a deformable part model is presented for discriminatively detecting objects in cluttered scenes. This method, however, assumes that the target object consists of a small number of deformable parts, which might not be able to model the subtle difference between similar image categories.

Our feature is similar in spirit to [3], though independently developed. We differ from [3] in that our features are automatically discovered instead of supervised by humans, making it a more scalable and convenient algorithm. Furthermore, we emphasize the dependence among image features, which is critical for demanding recognition tasks such as human and object interactions.

There has been a lot of work on discriminative feature selection [14, 16]. But most of the methods are not able to manage such a huge number of features (2 to the power of millions) as in the grouplets. Our algorithm is inspired by previous works [22, 29, 28] that also use data mining methods for feature selection. But compared to these previous methods, we take a step further to encode much more structured information in the feature representation.

7. Experiment

We first conduct experiments to analyze the properties of grouplets (Sec.7.1). The rest of this section then focuses on comparing using grouplets for human-object interaction classification and detection with a number of existing state-of-the-art methods. Apart from Sec.7.5, all experiments use the PPMI dataset introduced in Sec.2. In Sec.7.4 we use the original PPMI images. Data sets that are used from Sec.7.1 to 7.3 are obtained as follows. We first run a face detector [14] on all PPMI images. For each instrument, we manually select 200 detection results from PPMI+ and PPMI- images respectively. We then crop a rectangle region of the upper body of each selected detection result and normalize the region to 256×256 pixels so that the face size is 32×32.

7.1. Analysis of the Properties of the Grouplets

Effects of the grouplet size We use a 7-class classification task to analyze the properties of the mined grouplets (experiment details in Sec.7.2). Here we use an SVM with the histogram intersection kernel for classification.

Fig.6(left) shows the average distribution of different sizes of grouplets. Because the AND operation takes the smallest signal value of all feature units, it is unlikely that grouplets with a very large size can be mined. We observe that a majority of the mined grouplets contain 1, 2, or 3 feature units. Fig.6(right) shows the classification performance as the size of the grouplets increases. We see a big increase in accuracy using grouplets from size 1 to size 3. After this, the accuracy stabilizes even when including grouplets of bigger sizes. Two reasons might account for this observation: 1) the number of grouplets containing more than 3

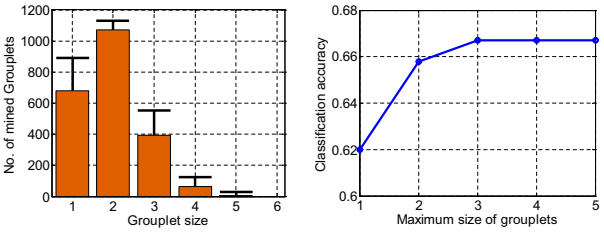

Figure 6. **left:** Average distribution of grouplets containing different number of feature units in the PPMI images. Error bars indicate the standard deviation to the mean among 7 activities. **right:** 7-class classification accuracy with respect to the number of feature units in included grouplets.

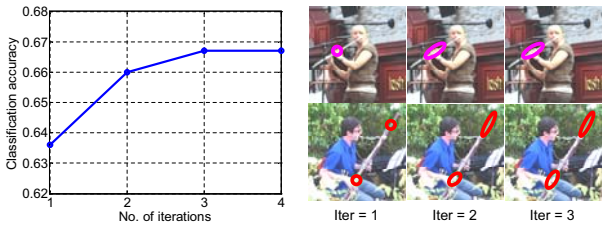

Figure 7. **left:** Classification accuracy with respect to the number of iterations (outer loop of Alg.1) of grouplet mining and parameter estimation. **right:** Spatial extent of grouplets mined in the 1st, 2nd, and 3rd iteration.

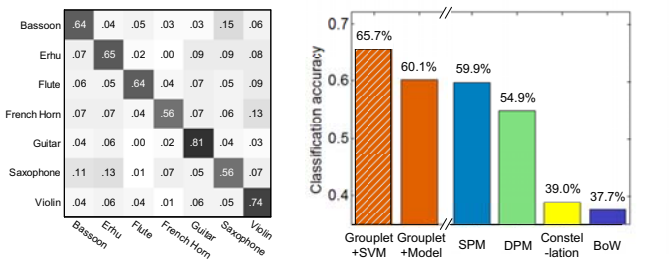

Figure 8. 7-class classification using the normalized PPMI+ images. **left:** Confusion matrix obtained by grouplet+SVM. The classification accuracy is 65.7%, whereas chance is 14%. **right:** Classification results of different methods: our grouplet+SVM, our grouplet+Model, a four-level spatial pyramid matching (SPM) [18], the deformable part model (DPM) [8], the constellation model [9], and bag-of-words (BoW). Y-axis indicates the classification accuracy of each method on the 7 classes.

feature units is small, and hence the overall contribution to classification is small; 2) much information in such grouplets is already contained in the grouplets of smaller sizes.

Effect of the Iterative Learning Procedure Given a set of training images, our algorithm iterates between a mining and a parameter estimation step. The idea is that each iteration offers a better refinement of the grouplet parameters (e.g. spatial extent of codeword), hence of the overall dscriminability. Fig.7(left) shows that the classification accuracy increases with respect to the iteration number. We observe the biggest gain between the first and the second iteration, indicating that with only two iterations, the method can obtain a good estimation of the spatial extent of each grouplet. Fig.7(right) shows that the estimation of the spatial extent of the grouplets align better with the visual features as the iteration increases, resulting in better grouplets.

7.2. Classification of Playing Different Instruments

Here we use our algorithm (grouplet+SVM and grouplet+Model, Sec.5) to classify images of people playing seven different musical instruments. For each class, 100 normalized PPMI+ images are randomly selected for training and the remaining 100 images for testing. We use three iterations of the iterative learning framework to mine around 2000 grouplets for each class. Fig.8(left) shows the confusion table obtained by grouplet+SVM with the histogram intersection kernel. We observe that the histogram intersection kernel performs better than the other kernels.

We compare our method with some other approaches. The results are shown in Fig.8(right). Both BoW and SPM [18] use the histogram representation, where BoW does not consider spatial information in image features while SPM accounts for some level of coarse spatial information by building histograms in different regions of the image. The BoW representation is followed by an SVM classifier with the histogram intersection kernel. Both DPM [8] and the constellation model [9] are part-based models, where DPM trains the classifier discriminatively and constellation model adopts a generative way.

We observe that our grouplet+SVM outperforms the other methods by a large margin. This suggests the effectiveness of the structural information in the mined grouplets. Furthermore, the method that combines grouplet with a generative model achieves comparable performance with SPM. This demonstrates that (1) the discriminatively mined grouplets carry the information that can distinguish images of different classes; (2) our parameter estimation step can effectively learn the weights of each mined grouplet.

7.3. Discriminating Playing from Not Playing

Our algorithm aims to learn discriminative structured information of human-object interactions. To demonstrate this, we conduct a classification experiment on PPMI+ vs. PPMI- datasets. For each instrument, we perform a binary classification task: whether the picture contains a person playing the instrument or a person not playing the instrument. Note that all images contain person(s) and instrument(s). The distinction between PPMI+ and PPMI- is only the way the person is interacting with the instrument.

We have 7 binary classification problems. In each problem, 100 normalized PPMI+ and 100 PPMI- images are randomly selected for training, and the other 200 images are used for testing. We mine around 4000 grouplets for both PPMI+ and PPMI- images of each instrument. In Table 1, our method is compared with the other approaches described in Sec.7.2. Due to space limitation, results of the constellation model, which performs on par with BoW, are

Instruments	SPM [18]	DPM [8]	BoW	Grouplet +Model	Grouplet +SVM
bassoon	71.5%	68.5%	64.5%	75.0%	**78.0%**
erhu	78.0%	75.5%	77.5%	**78.5%**	**78.5%**
flute	84.5%	79.0%	78.0%	85.0%	**90.5%**
French horn	78.5%	75.5%	71.5%	77.0%	**80.5%**
guitar	79.5%	**81.0%**	68.0%	73.0%	75.5%
saxophone	76.0%	76.5%	73.0%	75.0%	**78.5%**
violin	78.5%	75.5%	74.0%	83.5%	**85.0%**

Table 1. Classification results of PPMI+ (playing instrument) vs. PPMI- (co-occurring but not playing the instrument).

not listed in Table 1. We can see that our method outperforms the other methods on almost all the classes, especially on bassoon, flute, and violin, where our approach improves the accuracy by almost 10%. The only exception is guitar, where DPM achieves the best performance. The reason is that in the normalized images of people playing guitar, the guitar always occupies a big region at the left-bottom part of the image (Fig.10). Therefore it is not difficult for the part-based methods (DPM, SPM) to localize the guitar in each image. Fig.10(e) shows some PPMI- images with the grouplets that are mined for the corresponding PPMI+ images of the same instrument. Compared with Fig.10(d), much fewer grouplets are observed on PPMI- images.

7.4. Detecting Human and Object Interactions

Here, we test our approach's ability to detect activities in cluttered scenes. We use the original PPMI images as shown in Fig.2. In this experiment, 80 PPMI+ and 80 PPMI- randomly selected images of each instrument are used for training, and the remaining images for testing.

We first run a face detector on all images. We set a relatively low detection threshold to guarantee that almost all human faces are detected. Fig.9 shows that many false alarms occur after this step, at positions where no face is present or on a person who is not playing an instrument. Given each face detection, we crop out the neighboring region. Based on these regions, we mine the grouplets that are discriminative for detecting people playing each instrument. Then, an 8-class SVM classifier is trained to determine whether this detection contains a person playing one of the 7 instruments or not. This is a very challenging task (see Fig.9). The preliminary experiment result shows that, measured with area under the precision-recall curve, our algorithm significantly outperforms the SPM method [18]: we obtain a 45.7% performance, while SPM is 37.3%. We show examples of both successes and failures of our algorithm and SPM in Fig.9, from which we can see that SPM produces more false alarms than our method.

7.5. Result on Other Dataset - Caltech 101

Not only grouplets can be used for recognizing human-object interactions, but it is also a general framework to mine structured visual features in images. Therefore we also test our algorithm in an object recognition task using Caltech101 [6], in the same setting as in [11]. Table 2 compares our results with some previous methods. Other than the method in [10], our model performs on par with most of the state-of-the-art algorithms. It is important to note that this experiment is carried out without any additional tuning of the algorithm designed for activity classification. To accommodate objects that are not characterized by specific spatial structures (e.g. articulated animals), some design modifications should be applied to mine the grouplets.

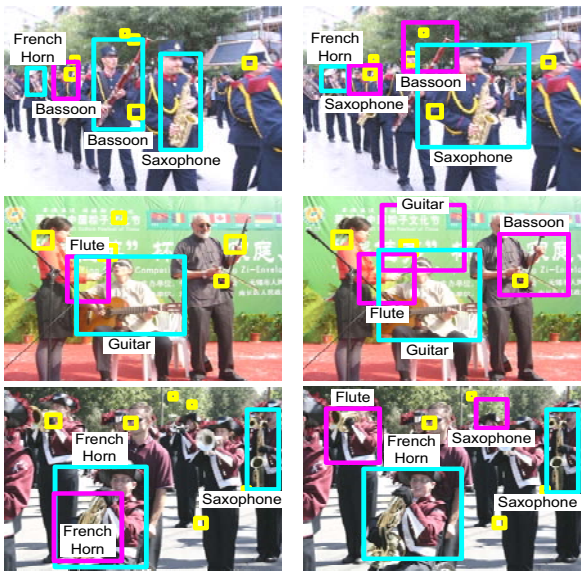

Figure 9. Examples of detection results by **(left)** our method and **(right)** SPM. Cyan and magenta rectangles denote the detection results and false alarms respectively. Bounding boxes in **(left)** are drawn by including all the grouplets that have large signals on the image region. Yellow rectangles show the face detection results which are classified as background.

Method	[2]	[11]	[30]	[10]	Grouplet+SVM
Accuracy	48%	59%	65%	77%	62%

Table 2. Recognition results on Caltech 101. The performance is measured by the average accuracy of the 101 classes.

8. Conclusion

In this work, we proposed a grouplet feature for recognizing human-object interactions. Grouplets encode detailed and structured information in the image data. A data mining method incorporated with a parameter estimation step is applied to mine the discriminative grouplets. One future research direction would be to link the mined grouplets with semantic meanings in the images to obtain deeper understanding of the scenes of human-object interactions.

Acknowledgement. This research is partially supported by an NSF CAREER grant (IIS-0845230), a Google research award, and a Microsoft Research Fellowship to L.F-F. We also would like to thank Juan Carlos Niebles and Jia Deng for helpful discussions.

References

[1] R. Agrawal and R. Srikant. Fast algorithms for mining association rules. In *VLDB*, 1994. 4, 9

[2] A. Berg, T. Berg, and J. Malik. Shape matching and object recognition using low distortion correspondences. In *CVPR*, 2005. 7

[3] L. Bourdev and J. Malik. Poselets: body part detectors trained using 3D human pose annotations. In *ICCV*, 2009. 5

[4] Y. Chen, L. Zhu, C. Lin, A. Yuille, and H. Zhang. Rapid inference on a novel AND/OR graph for object detection, segmentation and parsing. In *NIPS*, 2007. 2

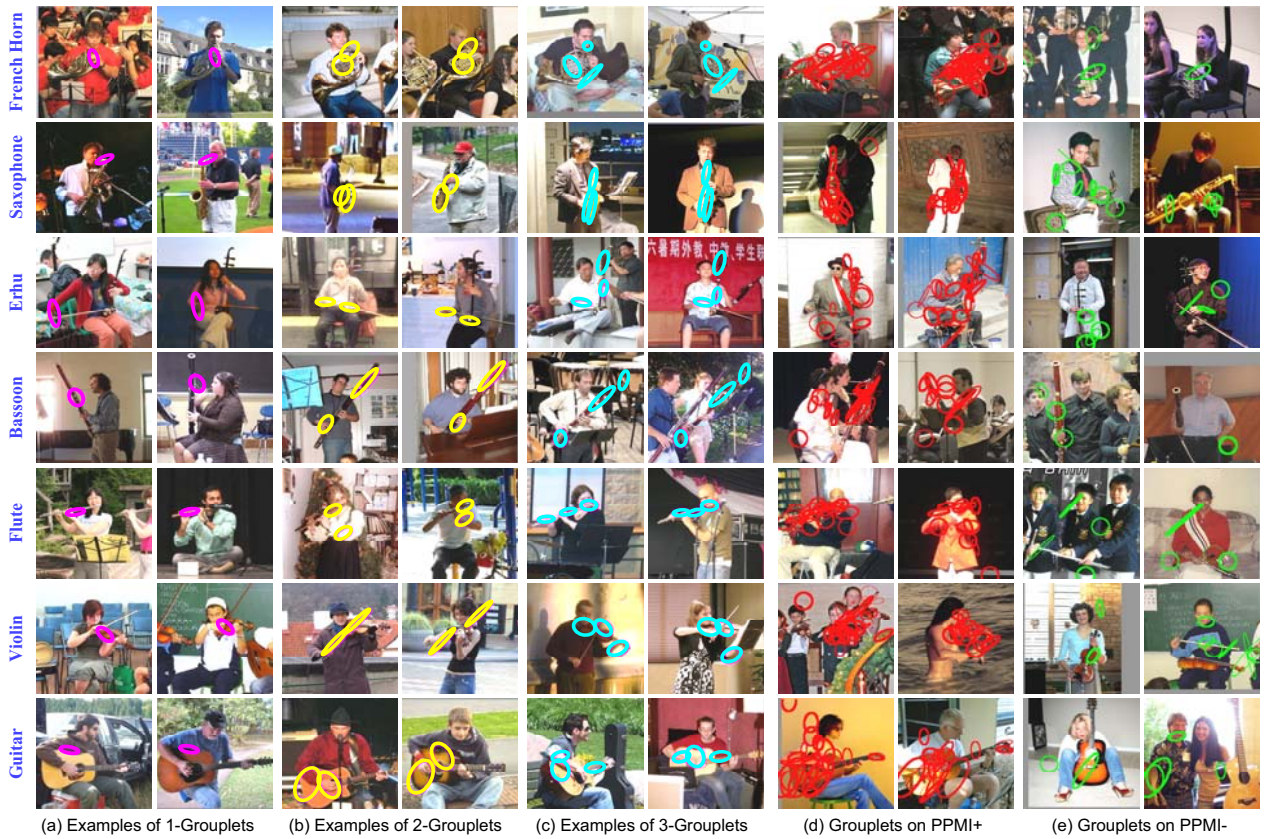

Figure 10. More example images and grouplets. **(a,b,c)** We show examples of 1, 2, and 3-grouplets on each of the two images for each class respectively. **(d,e)** On each image, we show all grouplets selected by the algorithm for this class whose signal values are stronger than a threshold. We can see that PPMI- images have a much smaller number of grouplets with strong signals than PPMI+ images.

[5] M. Everingham, L. van Gool, C. Williams, J. Winn, and A. Zisserman. The PASCAL voc 2008 Results. 1, 2

[6] L. Fei-Fei, R. Fergus, and P. Perona. Learning generative visual models from few training examples: an incremental bayesian approach tested on 101 object categories. In *GMBV Workshop*, 2004. 2, 7

[7] P. Felzenszwalb and D. Huttenlocher. Pictorial structures for object recognition. *Int. J. Comput. Vision*, 2005. 5

[8] P. Felzenszwalb, D. McAllester, and D. Ramanan. A discriminatively trained, multiscale, deformable part model. In *CVPR*, 2008. 5, 6

[9] R. Fergus, P. Perona, and A. Zisserman. Object class recognition by unsupervised scale-invariant learning. In *CVPR*, 2003. 2, 5, 6

[10] P. Gehler and S. Nowozin. On feature combination for multiclass object classification. In *ICCV*, 2009. 7

[11] K. Grauman and T. Darrell. The pyramid match kernel: discriminative classification with sets of image features. In *ICCV*, 2005. 7

[12] A. Gupta and L. Davis. Objects in action: an approach for combining action understanding and object perception. In *CVPR*, 2007. 2

[13] A. Gupta, A. Kembhavi, and L. Davis. Observing human-object interactions: Using spatial and functional compatibility for recognition. *IEEE T. Pattern Anal.*, 2009. 2

[14] C. Huang, H. Ai, Y. Li, and S. Lao. High performance rotation invariant multiview face detection. *IEEE T. Pattern Anal.*, 2007. 5

[15] M. Iacoboni and J. Mazziotta. Mirror neuron system: basic findings and clinical applications. *Ann. Neurol.*, 2007. 1

[16] L. Karlinsky, M. Dinerstein, and S. Ullman. Unsupervised feature optimization (UFO): simultaneous selection of multiple features with their detection parameters. In *CVPR*, 2009. 5

[17] J. Kim and I. Biederman. Where do objects become scenes? *J. Vision*, 2009. 1, 2

[18] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *CVPR*, 2006. 2, 3, 5, 6, 7

[19] D. G. Lowe. Object recognition from local scale-invariant features. In *ICCV*, 1999. 3

[20] J. C. Niebles and L. Fei-Fei. A hierarchical model of shape and appearance for human action classification. In *CVPR*, 2007. 5

[21] A. Oliva and A. Torralba. Modeling the shape of the scene: a holistic representation of the shape envelope. *Int. J. Comput. Vision*, 2001. 2

[22] T. Quack, V. Ferrari, B. Leibe, and L. van Gool. Efficient mining of frequent and distinctive feature configurations. In *ICCV*, 2007. 5

[23] D. Ramanan. Learning to parse images of articulated objects. In *NIPS*, 2006. 1

[24] S. Savarese, J. Winn, and A. Criminisi. Discriminative object class models of appearance and shape by correlations. In *CVPR*, 2006. 5

[25] J. Sivic, B. Russell, A. Efros, A. Zisserman, and W. Freeman. Discovering objects and their location in images. In *ICCV*, 2005. 5

[26] T. Tuytelaars and K. Mikolajczyk. Local invariant feature detectors: a survey. *Found. Trend. Comp. Graph. Vision*, 2008. 5

[27] B. Yao and L. Fei-Fei. Modeling mutual context of object and human pose in human-object interaction activities. In *CVPR*, 2010. 1

[28] J. Yuan, J. Luo, and Y. Wu. Mining compositional features for boosting. In *CVPR*, 2008. 5

[29] J. Yuan, Y. Wu, and M. Yang. Discovery of collocation pattern: from visual words to visual phrases. In *CVPR*, 2007. 5

[30] H. Zhang, A. Berg, M. Maire, and J. Malik. SVM-KNN: discriminative nearest neighbor classification for visual category recognition. In *CVPR*, 2006. 7

A. Iterative Apriori Mining

In this section, we describe some implementation details of the iterative Apriori mining method that are not covered in detail in Sec.4.2.

A.1. Generating Candidate l-Grouplets

In Algorithm 1, we need to generate candidate l-grouplets based on the mined $(l-1)$-grouplets. We use the same method as that in [1], which contains two steps: the *join* step and the *prune* step. In the join step, we find all the pairs of mined $(l-1)$-grouplets that have $(l-2)$ overlapping feature units. Then each pair will be merged to form a candidate l-grouplet. In the prune step, a candidate l-grouplet will be deleted if any of its $(l-1)$-subsets is not in the mined $(l-1)$-grouplets.

Example We use integer variables to denote feature units, and let the mined 2-grouplets be $\{\{1\ 2\}, \{1\ 3\}, \{2\ 3\}, \{3\ 5\}\}$. After the join step, the set of candidate 3-grouplets will be $\{\{1\ 2\ 3\}, \{1\ 3\ 5\}, \{2\ 3\ 5\}\}$. The prune step will delete $\{1\ 3\ 5\}$ because one of its subset $\{1\ 5\}$ is not in the mined 2-grouplets. $\{2\ 3\ 5\}$ will also be deleted because $\{2\ 5\}$ is not in the mined 2-grouplets. Then the final candidate 3-grouplet will be only $\{1\ 2\ 3\}$.

A.2. Adjusting Thresholds T_{Supp} and T_{Conf}

The number of mined grouplets depends on the threshold values T_{Supp} and T_{Conf}. It also depends on the images of different classes. If the visual appearance of the images in one class is more consistent than that in other classes, then even with large T_{Supp} and T_{Conf} values, many grouplets might be mined for this class. In order to achieve good classification performance, we hope that the number of mined grouplets for all the classes can be approximately the same. This is achieved by using class-dependent threshold values.

For each class, we set the expected number of grouplets and the initial threshold values T_{Supp} and T_{Conf}. If we obtain more grouplets than the expectation, then increase the values of T_{Supp} and T_{Conf}, otherwise decrease them. We repeat the above step until approximately the same number of grouplets as the expectation are mined. In this paper the initial thresholds are $T_{Supp} = 0.005$, $T_{Conf} = 1.75$.

B. Parameter Estimation for Grouplets

B.1. Model Equation

From the mining step, we have obtained a set of grouplets $\{\Lambda^1, \cdots, \Lambda^M\}$. The objective of parameter estimation is to update the spatial extent of the feature units, and learn a class-dependent weight for each mined grouplet. The two goals are achieved by jointly training the grouplets mined for all the classes. Given an image $\mathcal{I}$ with a class label c, we compute the likelihood of $\mathcal{I}$ given the parameters $\boldsymbol{\theta}$, where $\boldsymbol{\theta}$ contains the class-dependent weight $\boldsymbol{\pi}$ of each grouplet and the covariance parameter $\boldsymbol{\sigma}$ that governs the spatial extent of each feature unit in each grouplet:

$$\begin{aligned} p(\mathcal{I}, c|\boldsymbol{\theta}) &= p(c|\boldsymbol{\theta}) p(\mathcal{I}|c, \boldsymbol{\theta}) \\ &= p(c|\boldsymbol{\theta}) \sum_{m=1}^{M} p(\mathcal{I}, \Lambda^m|c, \boldsymbol{\theta}) \\ &= \frac{1}{C} \sum_{m=1}^{M} \left[p(\mathcal{I}|\Lambda^m, \boldsymbol{\sigma}) p(\Lambda^m|c, \boldsymbol{\pi}) \right] \end{aligned} \tag{5}$$

We assume that the classes are uniformly distributed, and hence $p(c|\boldsymbol{\theta}) = \frac{1}{C}$ in Eq.5, where C is the number of classes.

$p(\mathcal{I}|\Lambda^m, \boldsymbol{\sigma})$ denotes the likelihood of $\mathcal{I}$ given Λ^m. We assume that $p(\mathcal{I}|\Lambda^m, \boldsymbol{\sigma}) \propto p(\Lambda^m|\mathcal{I}, \boldsymbol{\sigma})$, and use the signal value of Λ^m on $\mathcal{I}$ to approximately describe $p(\mathcal{I}|\Lambda^m, \boldsymbol{\sigma})$. Furthermore, we approximate Eq.1 by

$$v \approx p(A|a_h) \cdot \mathcal{N}(x_h|x, \boldsymbol{\sigma}) \tag{6}$$

where $\{a_h, x_h\} = \arg\max_{a', x'} p(A|a') \cdot \mathcal{N}(x'|x, \boldsymbol{\sigma})$. With this approximation, we can avoid computing marginalization within the "ln" operation in model learning.

$p(\Lambda^m|c, \boldsymbol{\pi})$ models the importance of Λ^m for class c. It is expressed as a multinomial distribution,

$$\begin{aligned} p(\Lambda^m|c, \boldsymbol{\theta}) &= \prod_{c'=1}^{C} \mathrm{Mult}(\Lambda^m|\boldsymbol{\pi}_{:,c'})^{\delta(c,c')} \\ &= \prod_{c'=1}^{C} (\boldsymbol{\pi}_{m,c'})^{\delta(c,c')} \end{aligned} \tag{7}$$

where $\delta(c, c')$ equals 1 if $c = c'$ and otherwise 0. From Eq.7 we can see that $\boldsymbol{\pi}$ is a $M \times C$ matrix.

B.2. Learning

We use an EM approach to estimate the model parameters $\boldsymbol{\theta} = \{\boldsymbol{\pi}, \boldsymbol{\sigma}\}$, which is as follows,

1. Choose an initial setting for the parameters $\boldsymbol{\theta}^{\text{old}}$.
2. **E step** Evaluate $p(\Lambda^m|\mathcal{I}, c, \boldsymbol{\theta}^{\text{old}})$ for each $\{\mathcal{I}, c\}$ and $m = 1, \cdots, M$.
3. **M step** Evaluate $\boldsymbol{\theta}^{\text{new}}$ by $\boldsymbol{\theta}^{\text{new}} = \arg\max_{\boldsymbol{\theta}} \mathcal{Q}(\boldsymbol{\theta}, \boldsymbol{\theta}^{\text{old}})$, where

$$\begin{aligned} &\mathcal{Q}(\boldsymbol{\theta}, \boldsymbol{\theta}^{\text{old}}) \\ &= \sum_{\mathcal{I}, c} \sum_{m=1}^{M} \left[p(\Lambda^m|\mathcal{I}, c, \boldsymbol{\theta}^{\text{old}}) \ln p(\mathcal{I}, c, \Lambda^m|\boldsymbol{\theta}) \right]. \end{aligned}$$

4. Check for convergence of the parameter values. If the convergence criterion is not satisfied, then let $\boldsymbol{\theta}^{\text{old}} \leftarrow \boldsymbol{\theta}^{\text{new}}$ and return to step 2.

B.2.1 E Step

The E step estimates the likelihood of Λ^m given an image $\{\mathcal{I}, c\}$ and the model parameters $\boldsymbol{\theta}^{\text{old}}$. It is computed by

$$p(\Lambda^m|\mathcal{I}, c, \boldsymbol{\theta}^{\text{old}}) = \frac{p(\Lambda^m, \mathcal{I}, c|\boldsymbol{\theta}^{\text{old}})}{p(\mathcal{I}, c|\boldsymbol{\theta}^{\text{old}})} \tag{8}$$
$$= \frac{p(c)p(\Lambda^m|c, \boldsymbol{\pi}^{\text{old}})p(\mathcal{I}|\Lambda^m, \boldsymbol{\sigma}^{\text{old}})}{\sum_{l=1}^{M}[p(c)p(\Lambda^l|c, \boldsymbol{\pi}^{\text{old}})p(\mathcal{I}|\Lambda^l, \boldsymbol{\sigma}^{\text{old}})]}$$

B.2.2 M Step - Evaluation of $\boldsymbol{\pi}$

When updating $\boldsymbol{\pi}$, we must take account to the constraint of multinomial distribution parameters:

$$\sum_{l=1}^{M} \boldsymbol{\pi}_{l,k} = 1, \quad \forall k = 1, \cdots, C \tag{9}$$

We introduce a Lagrange multiplier λ_k for each $\boldsymbol{\pi}_{:,k}$. Furthermore we use a regularization parameter β to guarantee that the grouplets are almost equally distributed in the set of multinomial distributions, so as to avoid the situation that the learned group weights bias to some specific classes. Then $\boldsymbol{\pi}_{m,k}$ is estimated according to,

$$\frac{\partial}{\partial \boldsymbol{\pi}_{m,k}}\Bigg[\mathcal{Q}(\boldsymbol{\theta}, \boldsymbol{\theta}^{\text{old}}) - \beta\sum_{l=1}^{M}\left(\sum_{c'=1}^{C}\boldsymbol{\pi}_{l,c'} - \frac{C}{M}\right)^2 + \lambda_k\left(\sum_{l=1}^{M}\boldsymbol{\pi}_{l,k} - 1\right)\Bigg]$$
$$= \sum_{\mathcal{I},c}\left[p(\Lambda^m|\mathcal{I}, c, \boldsymbol{\theta}^{\text{old}}) \cdot \frac{\partial}{\partial \boldsymbol{\pi}_{m,k}} \ln p(\Lambda^m|c, \boldsymbol{\theta})\right] - 2\beta\left(\sum_{c'=1}^{C}\boldsymbol{\pi}_{m,c'} - \frac{C}{M}\right) + \lambda_k$$
$$= \sum_{\mathcal{I},c=k}\left[\frac{1}{\boldsymbol{\pi}_{m,k}}p(\Lambda^m|\mathcal{I}, c, \boldsymbol{\theta}^{\text{old}})\right] - 2\beta\left(\sum_{c'=1}^{C}\boldsymbol{\pi}_{m,c'} - \frac{C}{M}\right) + \lambda_k$$
$$= 0 \tag{10}$$

$$\therefore \boldsymbol{\pi}_{m,k} \approx \frac{\sum_{\mathcal{I},c=k} p(\Lambda^m|\mathcal{I}, c, \boldsymbol{\theta}^{\text{old}})}{2\beta\left(\sum_{c'=1}^{C}\boldsymbol{\pi}_{m,c'}^{\text{old}} - \frac{C}{M}\right) - \lambda_k} \tag{11}$$

where λ_k can be solved by considering the constraint

$$\sum_{l=1}^{M} \frac{\sum_{\mathcal{I},c=k} p(\Lambda^l|\mathcal{I}, c, \boldsymbol{\theta}^{\text{old}})}{2\beta\left(\sum_{c'=1}^{C}\boldsymbol{\pi}_{m,c'}^{\text{old}} - \frac{C}{M}\right) - \lambda_k} = 1 \tag{12}$$

Note that Eq.12 might have multiple solutions for λ_k. But we are only interested in the solution that satisfies the following constraint so that $\boldsymbol{\pi}_{m,c'}$ in Eq.11 is positive and can be the parameter of a multinomial distribution.

$$\forall m, \quad \lambda_k < 2\beta\left(\sum_{c'=1}^{C}\boldsymbol{\pi}_{m,c'}^{\text{old}} - \frac{C}{M}\right) \tag{13}$$

We prefer a large β for better regularization ability. But if β is too large, then Eq.13 cannot be well satisfied. In practice, we start from $\beta = 10,000$, and reduce β by half until a valid λ_k is obtained.

B.2.3 M Step - Evaluation of $\boldsymbol{\sigma}$

As shown in Fig.4, for a codeword A_w we learn the same distribution $\boldsymbol{\sigma}_w$, no matter what feature unit A_w belongs to and where A_w appears in the image. In the M-step we estimate $\boldsymbol{\sigma}_w$ by

$$\frac{\partial \mathcal{Q}(\boldsymbol{\theta}, \boldsymbol{\theta}^{\text{old}})}{\partial \boldsymbol{\sigma}_w} = \frac{\partial}{\partial \boldsymbol{\sigma}_w}\sum_{\mathcal{I},c}\sum_{l=1}^{M}\left[p(\Lambda^l|\mathcal{I}, c, \boldsymbol{\theta}^{\text{old}}) \ln p(\mathcal{I}|\Lambda^l, \boldsymbol{\sigma}_w)\right]$$
$$= \sum_{\mathcal{I},c}\sum_{l=1}^{M}\Bigg[p(\Lambda^l|\mathcal{I}, c, \boldsymbol{\theta}^{\text{old}}) \cdot \sum_{A_w=A^{l,j}} \frac{\partial \ln(p(A^{l,j}|a_{h^{l,j}}) \cdot \mathcal{N}(x_{h^{l,j}}|x^{l,j}, \boldsymbol{\sigma}_w))}{\partial \boldsymbol{\sigma}_w}\Bigg]$$
$$= 0 \tag{14}$$

where $A^{l,j}$ and $x^{l,j}$ denote the visual codeword and image location of the j-th feature unit in the l-th grouplet respectively. $h^{l,j}$ is an index variable as in Eq.6. Denoting $x_{h^{l,j}} - x^{l,j} = \chi_{l,j}$, we have

$$\sum_{\mathcal{I},c}\sum_{l=1}^{M}\Bigg\{p(\Lambda^m|\mathcal{I}, c, \boldsymbol{\theta}^{\text{old}}) \cdot \sum_{A_w=A^{l,j}}\left[(\boldsymbol{\sigma}_w)^{-T} - (\boldsymbol{\sigma}_w)^{-T}\chi_{l,j}\chi_{l,j}^T(\boldsymbol{\sigma}_w)^{-T}\right]\Bigg\} = 0 \tag{15}$$

Therefore

$$\boldsymbol{\sigma}_w = \frac{\sum_{\mathcal{I},c}\sum_{l=1}^{M}\left[p(\Lambda^l|\mathcal{I}, c, \boldsymbol{\theta}^{\text{old}}) \cdot \sum_{A_w=A^{l,j}} \chi_{l,j}\chi_{l,j}^T\right]}{\sum_{\mathcal{I},c}\sum_{l=1}^{M}\left[p(\Lambda^l|\mathcal{I}, c, \boldsymbol{\theta}^{\text{old}}) \cdot \sum_{A_w=A^{l,j}} 1\right]} \tag{16}$$