Contextualizing Histogram

Bingbing Ni, Shuicheng Yan, Ashraf Kassim
Electrical and Computer Engineering
National University of Singapore
Singapore, 117576
{g0501096, eleyans, eleashra}@nus.edu.sg

Abstract

In this paper, we investigate how to incorporate spatial and/or temporal contextual information into classical histogram features with the aim of boosting visual classification performance. Firstly, we show that the stationary distribution derived from the normalized histogram-bin co-occurrence matrix characterizes the row sums of the original histogram-bin co-occurrence matrix. This underlying rationale of the histogram-bin co-occurrence features then motivates us to propose the concept of general contextualizing histogram *process, which encodes the spatial and/or temporal contexts as local homogeneity distributions and produces the so called* contextualized histograms *by convoluting these local homogeneity distributions with the histogram-bin index images/videos. Finally, the third and even higher order contextualized histograms are instantiated for encoding more complicated and informative spatial and/or temporal contextual information into histograms. We evaluate these proposed methods on face recognition and group activity classification problems, and the results demonstrate that the contextualized histograms significantly boost the visual classification performance.*

1. Introduction

Histogram representations, e.g., Color Histogram, Histogram of Local Binary Patterns [15], and Bag-of-Words based on SIFT features [11], have been widely used in computer vision and multimedia communities for visual recognition, content based image retrieval, and video content analysis. The inability of conventional histogram features to convey spatial and temporal contextual information, however, greatly limits their discriminating power. Layout histograms and multi-resolution histograms [7] are the pioneering attempts to incorporate spatial contextual information for improving the discriminating capability of the histogram features. Instead of the indirect use of spatial contextual information, coherence vector [2] and autocorrelogram [8] were proposed to encode local spatial contextual information directly into histograms. Recently, Li et al. [9] introduced the spatial co-occurrence matrix based Markov chain model to encode the intra-histogram-bin and inter-histogram-bin relationships into histograms, where the initial and stationary distributions of the Markov chain model are combined to form the so-called Markov stationary features.

This paper investigates how to more generally and effectively incorporate spatial and/or temporal contextual information into classical histogram features for boosting visual classification performance. The contributions are two-fold. Firstly, we theoretically prove that there exists an informative trivial stationary distribution for the Markov chain model with the transition matrix as the normalized spatial histogram-bin co-occurrence matrix. This trivial stationary distribution is a normalized vector, where each element is the row sum of the spatial histogram-bin co-occurrence matrix. This proof offers an explicit semantic explanation for the derived Markov stationary features, from which we derive the homogeneity-aware Markov stationary features for eliminating the inherent ambiguities of the Markov stationary features proposed in [9], by considering only the mutually distinct pairs in computing the spatial histogram-bin co-occurrence matrix, *i.e.*, the diagonal elements of the histogram-bin co-occurrence matrix are set to be zeros.

Based on the above-mentioned theoretic analysis, we propose the concept of general *contextualizing histogram* process, where the local contextual structure and histogram features characterize an image or video from two complementary aspects, namely, *style* and *content*. The local contextual structure describes the histogram-bin homogeneity distribution information within an area with certain shape, and the convolution of local contextual structure and histogram-bin index image/video leads to the so-called *contextualized histogram*. Based on this new concept, the ternary (with its temporal extensions) or even higher order contextualized histograms are presented for encoding more

978-1-4244-3991-1/09/$25.00 ©2009 IEEE

complicated and informative local contextual information into histograms, where the local contextual structures can be triangle, T-shape, or L-shape, rather than the conventional binary pixel-pair. The homogeneity-aware Markov stationary features and the proposed ternary contextualized histograms are evaluated on two visual classification problems, *i.e.*, face recognition and human group activity classification, and the experimental results show significant improvement in accuracy brought by the ternary contextualized histograms as well as the encouraging gain from the homogeneity-aware Markov stationary features.

2. Co-occurrence based Context Modeling

2.1. Markov Stationary Features Revisited

The Markov Stationary Features (MSF) [9] was recently proposed to characterize spatial co-occurrence of histogram patterns based on the Markov chain model, which is shown to be generally superior over the coherence vector and auto-correlogram by incorporating both intra-histogram-bin and inter-histogram-bin information for visual representation. Here, we give a brief introduction of MSF as follows.

A visual image or video is quantized into K histogram bins denoted as $\mathbf{S} = \{c_1, ..., c_K\}$, and the MSF is a feature representation that can characterize both intra-histogram-bin spatial information and inter-histogram-bin spatial information. The spatial co-occurrence matrix is defined as $\mathbf{C} = [c_{ij}] \in \mathbb{R}^{K\times K}$ with each element as

$$c_{ij} = \#(p_1^c = c_i, p_2^c = c_j \mid ||p_1 - p_2||_1 = d), \quad (1)$$

where p_1 and p_2 are a pair of neighboring pixels with ℓ^1 distance as d[1], the corresponding histogram bin indices are denoted as p_1^c and p_2^c, respectively, and the $\#$ means the number of pairs satisfying all the conditions listed in parentheses. Note that the matrix $\mathbf{C}$ is symmetric and nonnegative. The co-occurrence matrix can be interpreted from a statistical view [9], and the corresponding transition matrix derived from the spatial co-occurrence matrix is defined as $\mathbf{P} = [p_{ij}] \in \mathbb{R}^{K\times K}$, where

$$p_{ij} = \frac{c_{ij}}{\sum_{k=1}^{K} c_{ik}}. \quad (2)$$

The above definition of $\mathbf{P}$ satisfies the basic properties of a Markov chain, namely,

$$1.\ p_{ij} \geq 0, \forall\ c_i, c_j \in \mathbf{S}. \quad (3)$$

$$2.\ \sum_{j=1}^{K} p_{ij} = 1,\ i = 1, 2, \cdots, K. \quad (4)$$

[1]Note that in this work, $d = 1$ for the MSF and homogeneity-aware MSF as well as contextualized histogram introduced afterward as in [9].

This representation of the Markov transition matrix is of K^2 dimension and may not be robust. In [9], the initial distribution, namely, an approximate auto-correlogram (a row vector $\boldsymbol{\pi}^a$ consisting of the normalized diagonal elements of $\mathbf{C}$), and the stationary distribution of the Markov chain (a row vector $\boldsymbol{\pi}$) are combined to form a $2K$ dimensional representation, called Markov stationary features, *i.e.*, $[\boldsymbol{\pi}^a, \boldsymbol{\pi}]$. The stationary distribution of the transition matrix is a K-dimensional row vector, denoted as $\boldsymbol{\pi} = (\pi_1, \pi_2, ..., \pi_K)$, satisfying

$$\boldsymbol{\pi} = \boldsymbol{\pi}\mathbf{P}. \quad (5)$$

For a regular Markov chain [4], its stationary distribution could be directly obtained as the solution to Eqn. (5). However, for general cases when the chain is irregular [4], there exists no unique solution to Eqn. (5), and then the informative stationary distribution is often approximated as the row average of the matrix $\mathbf{A}_n = \frac{1}{n+1}(\mathbf{I} + \mathbf{P} + \mathbf{P}^2 + \mathbf{P}^3 + ... + \mathbf{P}^n)$, where n is a large integer and set to be 50 as in [9]. In next subsection, we theoretically prove that for both regular and irregular Markov chains, there exists an informative trivial solution with explicit semantic for every transition matrix derived from a spatial co-occurrence matrix $\mathbf{C}$.

2.2. Justification of Informative Trivial Solution

Theorem The distribution $\boldsymbol{\pi}$, defined as $\pi_i = \frac{\sum_j c_{ij}}{\sum_i \sum_j c_{ij}}$, is a trivial stationary distribution for a Markov chain with the transition matrix $\mathbf{P}$ defined in Eqn. (2), namely, $\boldsymbol{\pi} = \boldsymbol{\pi}\mathbf{P}$.

Proof: Substituting $\pi_i = \frac{\sum_j c_{ij}}{\sum_i \sum_j c_{ij}}$ into the right side of Eqn. (5), we obtain

$$(\boldsymbol{\pi}\mathbf{P})_i = \sum_k \pi_k \times p_{ki} \quad (6)$$

$$= \sum_k \frac{\sum_j c_{kj}}{\sum_i \sum_j c_{ij}} \times \frac{c_{ki}}{\sum_j c_{kj}} \quad (7)$$

$$= \frac{\sum_k c_{ki}}{\sum_i \sum_j c_{ij}} = \frac{\sum_k c_{ik}}{\sum_i \sum_j c_{ij}} = \pi_i, \quad (8)$$

where the third equation is based on the symmetric property, *i.e.*, $c_{ik} = c_{ki},\ \forall\ i, k$. This proves that $\boldsymbol{\pi}$ defined as $\pi_i = \frac{\sum_j c_{ij}}{\sum_i \sum_j c_{ij}}$ is a trivial stationary distribution. □

This trivial solution has explicit semantic, that is, π_i characterizes the total co-occurrence number, $\sum_j c_{ij}$, for the c_i histogram pattern/bin. Moreover, if we denote n_d as the number of pixels with ℓ^1 distance as d for each pixel (except for the boundary ones), we have

$$\sum_j c_{ij} \doteq \#(p^c = c_i) \times n_d, \quad (9)$$

$$\sum_i \sum_j c_{ij} \doteq \sum_i \#(p^c = c_i) \times n_d = N n_d, \quad (10)$$

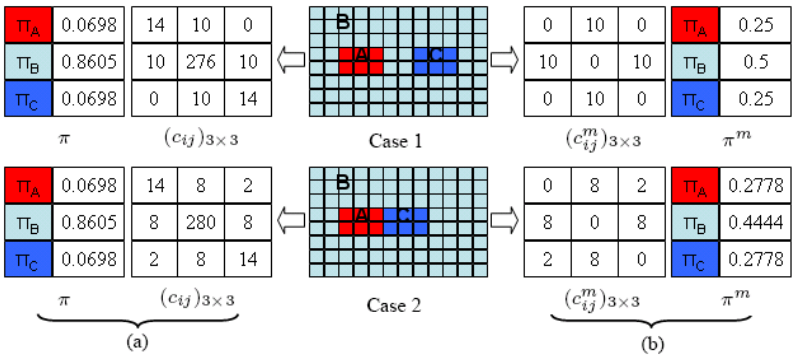


Figure 1. (a) An example which shows an informative trivial solution of MSF with no discriminant information. Case 1 and case 2 differ in both intra-histogram-bin and inter-histogram-bin relationships, however, their stationary distributions are the same for MSF. (b) An example where the homogeneity-aware MSF well characterizes the inter-histogram-bin spatial co-occurrence information. Note that we use d=1 for computing the spatial co-occurrence matrix. For better viewing, please see the color pdf file.

where p^c is the histogram bin index for a pixel p, and N is the total number of pixels for each image or video. Here, the $\doteq$ comes from the fact that the boundary pixels of an image may have fewer neighboring pixels with ℓ^1 distance as d. Then the semantic of the Markov stationary features can be further explained as

$$\pi_i \propto \sum_j c_{ij} \doteq \#(p^c = c_i) \times n_d \propto \#(p^c = c_i). \quad (11)$$

It means that the MSF described in [9] approximately equals to the original histogram features, and hence can only convey very limited spatial co-occurrence information. An illustrative example where MSF fails to convey discriminant information is shown in Figure 1(a). The success of MSF stems from its combination with the approximate auto-correlogram features $\boldsymbol{\pi}^a$, and the weighted difference between these two types of features implicitly characterizes inter-histogram-bin spatial co-occurrence information. This analysis also directly motivates us to propose a homogeneity-aware MSF in next subsection as the trivial stationary distribution of the transition matrix from the spatial co-occurrence matrix which only considers the inter-histogram-bin co-occurrence information.

2.3. Homogeneity-aware MSF

As proved above, the stationary distribution of the transition matrix approximately characterizes the row sums of the spatial co-occurrence matrix. Since the approximate auto-correlogram characterizes the intra-histogram-bin spatial co-occurrence information, it is desirable to obtain a complementary vector which characterizes the inter-histogram-bin spatial co-occurrence information only. In this section, we present a new MSF for such a purpose. First, a new spatial co-occurrence matrix $\mathbf{C}^m$ is defined as

$$c^m_{ij} = \#(p^c_1 = c_i, p^c_2 = c_j \mid ||p_1 - p_2||_1 = d),\ i \neq j, \quad (12)$$
$$c^m_{ii} = 0,\ i = 1, 2, ..., K. \quad (13)$$

The definition of p_{ij} is the same as in Eqn. (2), the resulting p_{ij} however is zero diagonal. Accordingly, the informative trivial solution to the Markov stationary distribution is $\boldsymbol{\pi}^m$ defined as

$$\pi^m_i = \frac{\sum_{j \neq i} c^m_{ij}}{\sum_i \sum_{j \neq i} c^m_{ij}}. \quad (14)$$

This new stationary distribution takes the homogeneity of the histogram bin pair into consideration, and hence its combination with the approximate auto-correlogram features is called homogeneity-aware Markov stationary features (HMSF), denoted as

$$\mathbf{x} = [\boldsymbol{\pi}^m, \boldsymbol{\pi}^a]. \quad (15)$$

The $\boldsymbol{\pi}^m$ and $\boldsymbol{\pi}^a$ characterize the inter-histogram-bin and intra-histogram-bin spatial co-occurrence information, respectively, and they are complementary to each other. Figure 1(b) shows the same example as in Figure 1(a) where original MSF conveys no discriminant information, but the homogeneity-aware MSF provides sufficient discriminant information for differentiating these two cases.

3. From HMSF to Contextualized Histogram

3.1. General Contextualizing Histogram Process

Beyond the statistical view of the homogeneity-aware MSF formulated by Markov chain model, the above theorem in Section-2.2 provides a more intuitive explanation on the rationale of homogeneity-aware MSF, that is, the auto-correlogram vector $\boldsymbol{\pi}^a$ describes the occurrence frequency

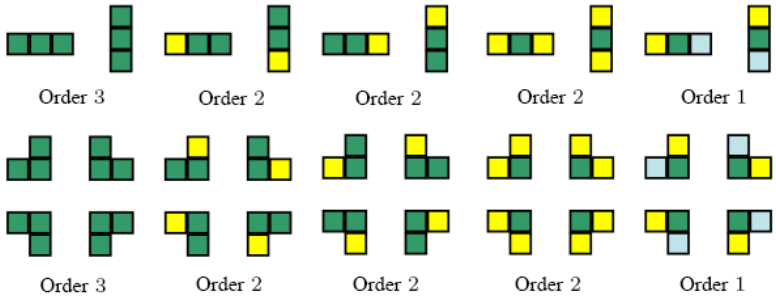


Figure 2. Illustration of the 30 ternary local contextual structures combining 5 homogeneities and 6 different shapes. Note that the order shown in the figure means the number of pixels belonging to the same histogram bin, and different colors represent different histogram bins.

of pixel pairs labeled with the same histogram bin, and $\boldsymbol{\pi}^m$ instead characterizes the occurrence frequency of the pixel pairs labeled with different histogram bins. Namely, HMSF utilizes two binary contextual structures, homogeneous and inhomogeneous pairs, to further decompose the original histograms into two sets of histograms with contextual information. The binary contextual structures are still limited in describing the contextual information, and more complicated local contextual structures, *e.g.*, triangle, T-shape, and L-shape, are desirable for characterizing more informative contextual information. The higher order contextual information is however far beyond the capability of Markov chain model, which is limited for characterizing binary relationship. For such a purpose, we present the concept of general *contextualizing histogram* process as below:

Definition (Contextualizing Histogram) For a histogram vector $\mathbf{h} = \{h_i\}_{i=1}^K$, the process of contextualizing histogram is to construct the so-called contextualized histogram $\mathbf{h}^s = \{h_i(s_j)\}_{i=1,j=1}^{K,M}$ based on a set of local contextual structures denoted as $\{s_j\}_{j=1}^M$, where $h_i(s_j)$ is the number of pixels belonging to the c_i histogram bin and with the local contextual structure as s_j, namely, the convolution of the local contextual structure s_j over the image/video with each pixel as the corresponding histogram bin index.

It is easy to verify that the homogeneity-aware MSF can be considered as the contextualized histogram based on two binary local contextual structures $\{s_1, s_2\}$, where the structure s_1 describes two homogeneous pixels with ℓ^1 distance as d and belonging to the same histogram bin, and the structure s_2 describes two inhomogeneous pixels with ℓ^1 distance as d and belonging to two distinct histogram bins. The general contextualizing histogram process also explains why the original MSF cannot sufficiently convey spatial co-occurrence information, since it is the contextualized histogram based on the local structure of two pixels with ℓ^1 distance as d yet without any constraints on the homogeneity. One advantage of the general contextualizing histogram process is that the length of the resulting histogram is only $K \times M$ (M=2 for binary contextual structures), not K^2 as in the case of directly using all the elements of the co-occurrence matrix $\mathbf{C}$. Another advantage of the general contextualizing histogram process is that it can be used for contextualizing histogram with local contextual structures of arbitrary orders as demonstrated afterwards.

3.2. Ternary Contextualized Histogram (TCH)

Here we introduce how to utilize the contextualizing histogram process for incorporating ternary spatial contextual information into general histogram features.

For a local contextual structure with only two pixels, there only exist two types of homogeneities, namely homogeneous and inhomogeneous as for the homogeneity-aware MSF. But for the ternary local contextual structure, there exist 5 types of local homogeneities, and the shape of the local structure may also change as shown in Figure 2. The ternary local contextual structure therefore encodes more complicated and informative local contextual information compared with the binary contextual structure, and in this work, we use 30 ternary contextual structures combining 5 types of homogeneities and 6 shapes. Denote these local structures as $\{s_j\}_{j=1}^{30}$, then for a histogram $\mathbf{h}$, its corresponding concatenated contextualized histogram is

$$\mathbf{h}^s = \{h_i(s_j)\} \in \mathbb{R}^{30K}, \tag{16}$$

where $h_i(s_j)$ means the number of pixels belonging to the c_i histogram bin and centering at which the local contextual structures belonging to the s_j category. Note again that the development of this proposed TCH is beyond the scope of Markov stationary feature since in the MSF framework, only the 2nd order relationship could be represented.

3.3. Temporal and Higher-order Extensions

The above contextualized histogram is defined in the spatial domain only, but for applications related with videos, temporal contextual information is also critical for characterizing the high-level semantics. We can also extend the general contextualizing histogram process into the

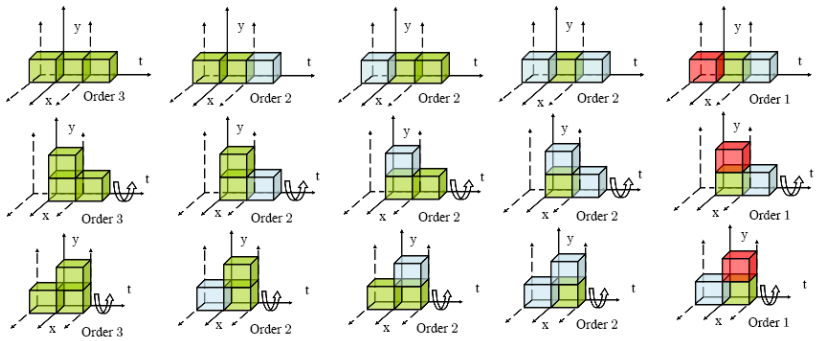


Figure 3. Illustration of the 15 local spatial-temporal contextual structures combining 5 types of homogeneities and 3 shapes. Note that the order shown under each contextual structure is for the homogeneity, and different colors represent different histogram bins.

temporal domain by exploiting both spatial and temporal configurations for local contextual structure. As shown in Figure 3, 15 types of local spatial and temporal contextual structures are defined based on 5 different types of homogeneities and 3 different shapes. Note that the rotation in the spatial domain is considered invariant for reducing the number of contextual structures, and the | shape structure appearing only within one frame is not considered due to the lack of temporal information. In this work, we utilize these 15 local spatial-temporal contextual structures to incorporate the contextual information into general histogram features, and the experiments on human group activity classification well justify the effectiveness of these spatial-temporal contextualized histogram features.

Beyond binary and ternary local contextual structures, the concept of local contextual structure can be naturally extended to the fourth or even higher orders. More types of homogeneities and shapes can be defined if higher order local contextual structures are used. It is predictable that better discriminating power may be gained but more computational cost and storage space are required for higher order contextualized histograms.

4. Related Works

Feature Descriptors vs. Contextualized Histograms The popular image descriptors, *e.g.*, SIFT [11] and Histograms of Oriented Gradients [5], also consider the image local spatial contextual information. The contextualized histogram is different from these descriptors in the following aspects. Firstly, the inputs to the general contextualizing histogram process are images/videos with each pixel quantized as index of a histogram bin rather than original intensity/color values for feature descriptors, and the feature descriptors cannot be directly applied on these histogram-bin index images/videos. Secondly, the approach in [5] to summarize certain quantized features within overlapping image cells is a general post-processing strategy, which can also be used to further enhance the performance of the proposed contextualized histograms. Finally, the proposed contextualizing histogram process is general, and is able to take the quantized SIFT and oriented gradient features as inputs to construct specialized contextualized histograms. The contextualized histograms based on SIFT features shall be further evaluated in the experimental section.

Relaxed Matching Kernels vs. Contextualized Histograms Recently, several histogram based kernels have been proposed. Grauman and Darrell proposed a pyramid matching kernel (PMK) [6] which represents the image by a set of histograms generated by recursively coarsening the bins/feature space partitions. To incorporate part of the spatial information, Bosch et al. later proposed a spatial pyramid matching kernel (SPMK) [3], where the original feature is augmented with a location descriptor and the pyramid is formed by coarsening the location component. Ling et al. proposed a method called proximity distribution kernel (PDK) [10] which adopts the concept of point pairs augmented with a relative distance measurement. The pyramid is constructed by gradually increasing the relative distance. Recently, Vedaldi and Soatto proposed a relaxed matching kernel (RMK) [19] to generalize the above kernel based representations, *e.g.*, PMK, SPMK, PDK, and also developed a new kernel called graph matching kernel (GMK). Our proposed general contextualizing histogram progress is essentially different from these kernels in terms of the underlying philosophy. More specifically, these kernels attempt to progressively merge the partitions/bins, while the contextualizing histogram progress aims to incorporate spatial and/or temporal contextual information into histograms by splitting bins instead of merging them.

5. Experiments

5.1. Data Sets

Two face databases CMU PIE [18] and FRGC V1.0 [16] are used in the face recognition experiments. We used 3329

Table 1. Summary on the BEHAVE human group activity database.

Activity category	Approach	Fight	InGroup	RunTogether	Split	WalkTogther	Total
No. of segments	19	12	25	6	17	6	85

Table 2. A summary of the recognition rates (%) for face classification on the FRGC V1.0 and CMU PIE databases.

FRGC V1.0 Dataset						CMU PIE Dataset					
Feature	Gray Level	LBP	DOG	SIFT-32	SIFT-128	Feature	Gray Level	LBP	DOG	SIFT-32	SIFT-128
Image Size: 100×100						Image Size: 64×64					
Histogram	35.28	39.41	27.43	54.00	68.50	Histogram	44.32	74.09	67.59	64.00	78.50
MSF	36.80	39.73	34.61	54.00	69.00	MSF	46.35	77.96	74.65	62.00	76.00
HMSF	40.64	41.96	37.22	57.25	69.25	HMSF	49.66	78.39	79.44	65.50	77.75
TCH	**47.19**	**46.62**	**44.40**	**62.75**	**72.50**	TCH	**57.58**	**82.57**	**86.68**	**70.25**	**78.75**
Image Size: 50×50						Image Size: 32×32					
Histogram	33.79	58.11	30.36	54.00	60.75	Histogram	42.66	77.47	63.72	52.00	55.75
MSF	35.52	59.67	37.33	53.00	60.50	MSF	44.57	81.22	70.23	48.50	53.75
HMSF	38.88	62.88	42.95	54.75	61.75	HMSF	46.29	81.71	72.13	52.25	55.75
TCH	**49.03**	**69.25**	**58.68**	**62.25**	**64.50**	TCH	**55.37**	**90.42**	**84.16**	**55.75**	**56.25**
Image Size: 25×25						Image Size: 16×16					
Histogram	28.74	55.28	22.16	35.50	37.00	Histogram	38.24	73.30	55.13	35.25	39.75
MSF	30.12	58.71	30.58	35.00	38.50	MSF	39.04	75.20	62.86	35.00	40.25
HMSF	32.98	60.20	36.20	38.50	38.25	HMSF	39.10	73.71	64.76	36.50	38.50
TCH	**47.15**	**67.23**	**56.42**	**46.50**	**40.75**	TCH	**51.44**	**82.93**	**78.58**	**37.75**	**42.75**
Image Pyramid						Image Pyramid					
Histogram	36.51	69.53	46.62	60.25	59.00	Histogram	45.61	87.91	78.39	49.75	56.25
MSF	36.76	71.37	53.98	60.50	58.75	MSF	47.64	89.44	81.52	48.50	55.00
HMSF	40.37	72.36	57.72	59.50	59.50	HMSF	48.80	88.83	83.00	50.00	55.25
TCH	**52.24**	**76.32**	**69.71**	**65.00**	**61.50**	TCH	**58.50**	**92.45**	**89.69**	**52.50**	**57.75**

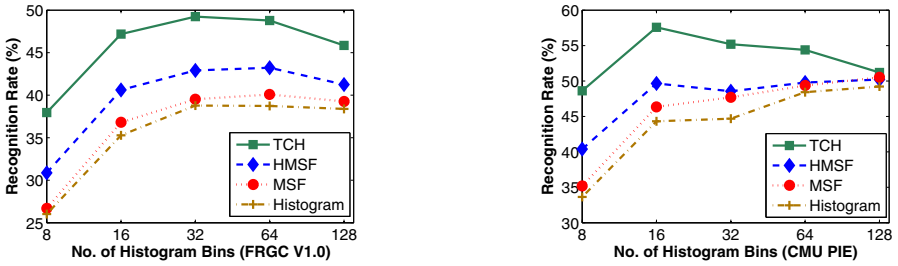
and 5658 frontal face images from 68 and 275 individuals with varying expressions and illuminations for these two databases. For CMU PIE database, the subset from the pose indexed as C27 is used. The images are of gray scale and with size as 64×64 and 100×100 pixels, respectively. The databases are randomly split into equal parts for training and testing.

The experiment on human group activity classification is based on the labeled video sequence in the *BEHAVE* human behavior database [20]. This video sequence contains different group activities recorded in an outdoor park scene at a frame rate of $25 fps$ and with image resolution of 640×480 pixels. Note that there may exist multiple labels for a single video segment, we manually select only one label which gives an overall description to the video segment, *e.g.*, if activity *WalkTogehter* and *Approah* exist at the same time, we take *Approach* as the label. The detailed information on these video segments is listed in Table 1. We use leave-one-out scheme for this experiment because of the limited number of samples in the database.

5.2. Face Recognition based on Ternary Contextualized Histograms

To evaluate the general performance of the homogeneity-aware MSF and the ternary contextualized histogram features, we exploit four different low-level features widely used for visual classification, namely, original gray level, local binary pattern (LBP) [15], direction of gradient (DOG) [5], and SIFT [11]. For gray level features, 16 histogram bins are used that correspond to different levels of image intensities. For LBP, we used the uniform LBP features which lead to 59 bins for histogram quantization. For direction of gradient, we use 16 bins that uniformly divide the whole direction space into 16 intervals. For SIFT, we calculate the feature vector at every pixel location and with fixed scale as well as direction, tuned to be optimal in performance. The original 128-dimensional descriptor is projected to 40-dimensional by Principal Component Analysis [16] for computational efficiency. We generate the code book by K-Means using the training images and after that all the images are encoded using this dictionary. We evaluated the algorithm with a dictionary of either 32 visual words (SIFT-32) or 128 visual words (SIFT-128, in a size similar to [10]). We found that the performance of SIFT will decrease if the dictionary size is further increased. One possible explanation of this phenomena is because of the small size and well aligning of the face image. Also, for the experiments on SIFT features, the size of both training and testing data sets are reduced to 800 images respectively for avoiding large computational burden. We vary the size of the input image by down-sampling with bilinear interpolation to validate the robustness of the representations, and also a pyramid scheme is used to evaluated the possibility to further boost performance. We do not further compare the proposed structure histograms with coherence vector and auto-correlogram features, since the MSF shows to be superior over them [9].

Although many stronger classification algorithms, *e.g.*, Support Vector Machine (SVM) [12], exist for further improving classification accuracy, in this work, we use the simple nearest neighbor classifier for final classification to better identify the gap between different histogram features and avoid the effect of the consequent strong classifiers, e.g., SVM. The dissimilarity measurement is based on χ^2



(a) (b)

Figure 4. Comparison robustness analysis with different histogram bin numbers. (a) Recognition rate vs. histogram bin number (FRGC V1.0) at image size of 100×100 pixels and with gray level features. (b) Recognition rate vs. histogram bin number (CMU PIE) at image size of 64×64 pixels and with gray level features.

distance between two histogram vectors $\mathbf{x}$ and $\mathbf{y}$, namely,

$$D(\mathbf{x}, \mathbf{y}) = \frac{1}{2} \sum_j \frac{(x_j - y_j)^2}{x_j + y_j}. \tag{17}$$

The comparison results of original histogram, MSF, homogeneity-aware MSF and TCH are listed in Table 2, from which the following observations can be made: 1) the original histogram generally gives the lowest recognition rates, and MSF improves the performance as reported in [9]; 2) the homogeneity-aware MSF generally gives higher performance than the original MSF; 3) the proposed ternary contextualized histogram features significantly outperform all the other histogram representations; and 4) the image pyramid scheme can further improve the performances for all these features, and the pyramid version of TCH is the best for all the evaluated features.

We experimentally evaluate the robustness of contextualized histograms with respect to the change of histogram bin number. The comparison results based on gray level features with original image size are shown in Figure 4, and we can see that, when the number of histogram bins varies, the ternary contextualized histogram consistently outperforms all the other histogram representations. Note that the case for different histogram features to be similar in performance (CMU PIE database with bin number as 128) is caused by the insufficient number of pixels for computing statistically robust contextualized histograms.

We also compared MSF, HMSF and TCH with the relaxed matching kernels, *i.e.*, PMK, PDK and GMK. Note that we do not compare these features with SPMK since its block strategy can also be utilized for further enhancing the performances of all these features. For PMK and GMK, we start from the finest partition/bin quantization, *e.g.*, 16 bins for gray level and DOG, 59 bins for LBP and 32/128 bins for SIFT, and recursively merge two bins into one. For PMK, PDK and GMK, we quantize the spatial distance into 8 layers/levels. Note that the bin layer number and distance layer number we used are adjusted to be optimal for these kernel based methods. We use χ^2 and L_1 distances for dissimilarity measurements. For PMK and GMK, the bin quantization levels are weighted in a way as proposed in [6, 19]. The nearest neighbor method is used for classification. The comparison results are shown in Table 3. From the results, we can observe that 1) MSF and HMSF generally can achieve higher accuracies than PMK, since PMK does not explicitly encode spatial contextual information; 2) PDK and GMK outperforms MSF and HMSF generally owing to their involved more redundant co-occurrence information than the PMK; and 3) the proposed TCH generally achieves the best accuracy although its descriptor size is much smaller than PDK and GMK.

5.3. Group Activity Classification by Ternary Temporal Contextualized Histograms

There exist methods for human group activity classification [17], but most of these are based on the assumption that each human in the scene has been tracked as a bounding box [13, 14]. However, robust tracking of objects itself is difficult especially when the scene is crowded and image quality is poor. Similar to [1], we use optical flows as raw features. For given video inputs, optical flows are extracted on down-sampled images of size 320×240 pixels, and then a 5×5 median filter is applied to eliminate the noises. Both the magnitudes and the directions of the optical flows are quantized into 8 bins, respectively. After this processing, the temporal extension of the ternary contextualized histogram features is used for extracting features to represent the videos. We also down-sampled the frame rate and the spatial image size to evaluate algorithmic robustness. The χ^2 distance is combined with the nearest neighbor classifier for final classification. The comparison classification results are listed in Table 4, from which we can observe that the temporal ternary contextualized histogram features boost the accuracy significantly compared with all other evaluated histogram features.

6. Conclusions

In this paper, we proved that there exists a trivial informative solution for the stationary distribution of the Markov chain model with the transition matrix as the normalized spatial histogram-bin co-occurrence matrix. This new insight motivated the development of the homogeneity-

Table 3. Comparison recognition rates of MSF, HMSF and TCH vs. relaxed matching kernels, *i.e.*, PMK, PDK and GMK on FRGC V1.0 and CMU PIE databases. Note that the results for relaxed matching kernels based on L_1 distance metric are listed in the parentheses.

	Feature Size	Gray Level	LBP	DOG	SIFT-32	SIFT-128
FRGC V1.0 Dataset, Image Size: 50×50						
Histogram	K	33.79	58.11	30.36	54.00	60.75
MSF	$2K$	35.52	59.67	37.33	53.00	60.50
HMSF	$2K$	38.88	62.88	42.95	54.75	61.75
PMK	$\approx 2K$	33.33(25.56)	58.32(51.04)	31.14(28.42)	52.75(51.75)	61.50(59.75)
PDK	MK^2	42.31(36.44)	**72.43**(71.16)	55.81(54.83)	61.00(59.25)	64.25(62.75)
GMK	$\approx \frac{4}{3}MK^2$	41.39(35.03)	70.80(68.15)	55.36(53.48)	60.00(59.00)	63.50(61.50)
TCH	$30K$	**49.03**	69.25	**58.68**	**62.25**	**64.50**
CMU PIE Dataset, Image Size: 32×32						
Histogram	K	42.66	77.47	63.72	52.00	55.75
MSF	$2K$	44.47	81.22	70.23	48.50	53.75
HMSF	$2K$	46.29	81.71	72.13	52.25	55.75
PMK	$\approx 2K$	39.90(34.99)	78.51(75.38)	64.09(60.10)	52.50(50.50)	**56.25**(56.00)
PDK	MK^2	45.18(40.02)	81.95(81.46)	71.21(69.86)	50.75(50.00)	55.50(55.00)
GMK	$\approx \frac{4}{3}MK^2$	42.85(36.83)	81.71(80.91)	72.56(71.03)	50.00(49.25)	55.00(54.50)
TCH	$30K$	**55.37**	**90.42**	**84.16**	**55.75**	**56.25**

Table 4. A summary of the leave-one-out accuracies (%) for human group activity classification on BEHAVE database. Note that *TTCH* means the ternary temporal contextualized histograms.

Frame Rate	12.5 fps			6.25 fps		
Image Size	320×240	160×120	80×60	320×240	160×120	80×60
Histogram	48.24	48.24	47.60	42.35	42.35	43.53
MSF	48.24	48.24	50.59	45.88	47.60	47.60
HMSF	48.24	48.24	48.24	48.24	49.41	50.59
TTCH	**54.12**	**52.94**	**52.94**	**56.47**	**56.47**	**56.47**

	Image Pyramid	Image Pyramid
Histogram	48.96	44.25
MSF	51.31	48.32
HMSF	48.24	51.31
TTCH	**54.12**	**57.65**

aware Markov stationary features and the concept of general contextualizing histogram process for encoding the spatial and/or temporal contextual information into general histogram features. The ternary contextualized histogram and its temporal extensions derived from the third order contextualizing histogram process are shown to be very effective in both face recognition and human activity classification problems, and greatly outperform the state-of-the-art attempts to encode contextual information into histograms.

Acknowledgment

This work is supported by NRF/IDM Program, under research Grant NRF2008IDM-IDM004-029.

References

[1] E. Andrade, S. Blunsden, and R. Fisher, Hidden Markov Models for Optical Flow Analysis in Crowds, *ICPR*, pp. 460-463, 2006.

[2] S. Birchfield and S. Rangarajan, Spatiograms Versus Histograms for Region-based Tracking, *CVPR*, pp. 1158-1163, 2005.

[3] A. Bosch, A. Zisserman, X. Munoz, Representing Shape with a Spatial Pyramid Kernel, *Proceedings of the 6th ACM international Conference on Image and Video Retrieval*, pp. 401 - 408, 2007.

[4] L. Breiman, Probability, *Society for Industrial and Applied Mathematics*, Chapter 7, 1992.

[5] N. Dalai and B. Triggs, Histograms of Oriented Gradients for Human Detection, *CVPR*, pp. 886-893, 2005.

[6] K. Grauman and T. Darrell, The Pyramid Match Kernel: Discriminative Classification with Sets of Image Features, *ICCV*, vol. 2, pp. 1458-1465, 2005.

[7] E. Hadjidemetriou, M. Grossberg, and B. Nayar, Multiresolution Histograms and Their Use for Recognition, *TPAMI*, 26(7):831-847, 2004.

[8] J. Huang, S. Kumar, M. Mitra, W. Zhu, and R. Zabih, Spatial Color Indexing and Applications, *IJCV*, 35(3):245-268, 1999.

[9] J. Li, W. Wu, T. Wang, and Y. Zhang, One Step Beyond Histogram: Image Representation using Markov Stationary Features, *CVPR*, 2008.

[10] H. Ling and S. Soatto, Proximity Distribution Kernels for Geometric Context in Category Recognition, *CVPR*, 2007.

[11] D. Lowe, Distinctive Image Features from Scale-invariant Keypoints, *International Journal of Computer Vision*, 60(2):91-110, 2004.

[12] D. Meyer, F. Leisch, and K. Hornik, The Support Vector Machine under Test, *Neurocomputing*, 55(1):169-186, 2003.

[13] J. Nascimento, M. Figueiredo, and J. Marques, Recognition of Human Activities using Space Dependent Switched Dynamical Models, *ICIP*, 2005.

[14] J. Nascimento, M. Figueiredo, and J. Marques, Segmentation and Classification of Human Activities, *Proceedings of International Workshop on Human Activity Recognition and Modelling*, 2005.

[15] T. Ojala, M. Pietikainen, and T. Maenpaa, Multiresolution Gray-scale and Rotation Invariant Texture Classification with Local Binary Patterns, *TPMAI*, 24(7):971-987, 2002.

[16] P. Phillips P. Flynn, T. Scruggs, K. Bowyer, J. Chang, K. Hoffman, J. Marques, J. Min, and W. Worek, Overview of the Face Recognition Grand Challenge, *CVPR*, pp. 947–954, 2005.

[17] P. Ribeiro and J. Victor, Human Activity Recognition from Video: Modeling, Feature Selection and Classification Architecture, *Proceedings of International Workshop on Human Activity Recognition and Modelling*, 2005.

[18] T. Sim, S. Baker, and M. Bsat, The CMU Pose, Illumination, and Expression Database, *TPAMI*, 25(2):1615-1618, 2003.

[19] A. Vedaldi and S. Soatto, Relaxed Matching Kernels for Robust Image Comparison, *CVPR*, 2008.

[20] The BEHAVE Website: *http://homepages.inf.ed.ac.uk/rbf/BEHAVE/*.