DIRECTED MARKOV STATIONARY FEATURES FOR VISUAL CLASSIFICATION

Bingbing Ni, Shuicheng Yan, and Ashraf Kassim

Department of Electrical and Computer Engineering, National University of Singapore

ABSTRACT

We investigate how to effectively incorporate spatial structure information into histogram features for boosting visual classification performance motivated by recently proposed Markov Stationary Features (MSF). First, we show that due to the symmetric property of the image occurrence modeling procedure, the stationary distribution derived from the normalized co-occurrence matrix has a trivial informative solution which only approximates the original histogram representation, i.e., does not encode proper spatial structure information. To eliminate this ambiguity, we propose in this work the so called Directed Markov Stationary Features (DMSF) to encode spatial information into histogram features, and the asymmetric essence of the co-occurrence matrices in DMSF avoids the trivial informative solutions in MSF. Extensive experiments on face recognition show the significant performance improvement brought by our proposed DMSF.

Index Terms— Visual classification, Markov Stationary Features, Directed Markov Stationary Features

1. INTRODUCTION

Histogram representations are widely used in computer vision and multimedia communities for visual classification, content based image retrieval, and video content analysis. The inability of conventional histogram features to convey spatial structure information, however, greatly limits their discriminating power. Layout histograms and multi-resolution histograms [1] are the pioneering attempts to incorporate spatial structure information for improving the discriminating capability of histogram features. Instead of the indirect use of spatial information, coherence vector [2] and auto-correlogram [3] were proposed to encode local spatial structure information directly into histograms. Recently, Li et al. [4] introduced the spatial co-occurrence matrix based Markov chain model to encode the intra-bin and inter-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 (MSF).

For MSF, we prove in this paper that there exists an informative trivial solution for the stationary distribution, given that the co-occurrence modeling of the image is symmetric, i.e., the co-occurrence matrix is symmetric. Under this condition, we show that the trivial solution is the normalized vector, where each element corresponds to the row sum of the spatial co-occurrence matrix. It means that this trivial solution only approximates the original histogram representation, and cannot effectively encode the spatial information. To eliminate this ambiguity as well as to boost the visual classification performance, we propose to compute the so called directed co-occurrence matrices, and a set of MSFs are calculated based on these asymmetric matrices. This new encoding framework is referred to as *Directed Markov Stationary Features* (DMSF). The direct advantages of DMSF over MSF include: 1) It avoids the trivial solution in MSF; 2) It effectively encodes the directional spatial context information.

The rest of this paper is organized as follows. Section 2 revisits the previous proposed Markov Stationary Features and theoretically shows its inherent drawback. In Section 3 we present the Directed Markov Stationary Features to address the problem associated with MSF. Section 4 demonstrates the experimental results on face recognition problem and Section 5 concludes this paper.

2. SPATIAL CO-OCCURRENCE BASED HISTOGRAMS

2.1. Markov Stationary Features Revisited

The Markov Stationary Features (MSF) [4] was recently proposed to characterize spatial co-occurrence of histogram patterns based on Markov Chain models, which is shown to be generally superior over the coherence vector and auto-correlogram by incorporating both intra-bin and inter-bin co-occurrence information for visual representation. Here, we give a brief introduction of MSF as follows.

The visual image or video is quantized into K histogram bins $\mathbf{S} = \{c_1, ..., c_K\}$, 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

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

where p_1 and p_2 are a pair of neighboring pixels with distance not larger than d (d is set as $\frac{\sqrt{2}}{2}$ in this work), the corresponding 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 brackets. Note that the matrix $\mathbf{C}$ is symmetric and nonnegative. The co-occurrence matrix can be interpreted from a statistical view [4], 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

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

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

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

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

This representation of the Markov transition matrix is of K^2 dimension and may not be robust. In [4], the initial distribution, namely, the auto-correlogram (a row vector $\boldsymbol{\pi}_a$), 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

$$\boldsymbol{\pi} = \boldsymbol{\pi}\mathbf{P}. \tag{5}$$

For a regular Markov chain [5], the stationary distribution could be directly obtained as the solution to Eqn. (5). However, for general cases when the chain is irregular [5], there exists no unique solution to Eqn. (5), and then the informative stationary distribution is often approximated as the row average of the matrix

$$\mathbf{A}_n = \frac{1}{n+1}(\mathbf{I} + \mathbf{P} + \mathbf{P}^2 + \mathbf{P}^3 + ... + \mathbf{P}^n), \tag{6}$$

where n is a large integer. In next subsection, we prove that there exists an informative trivial solution with explicit semantics for every transition matrix derived from a spatial co-occurrence matrix.

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 solution to 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

$$\begin{aligned} (\boldsymbol{\pi}\mathbf{P})_i &= \sum_k \pi_k \times p_{ki} & (7) \\ &= \sum_k \frac{\sum_j c_{kj}}{\sum_i \sum_j c_{ij}} \times \frac{c_{ki}}{\sum_j c_{kj}} & (8) \\ &= \frac{\sum_k c_{ki}}{\sum_i \sum_j c_{ij}} = \frac{\sum_k c_{ik}}{\sum_i \sum_j c_{ij}} = \pi_i, & (9) \end{aligned}$$

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

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. Moreover, if we denote n_d as the number of pixels with ℓ^2 distances not larger than d for each pixel (except for the boundary ones), we have

$$\begin{aligned} \sum_j c_{ij} &\doteq \#(p^c = c_i) \times n_d, & (10) \\ \sum_i \sum_j c_{ij} &\doteq \sum_i \#(p^c = c_i) \times n_d = N \times n_d, & (11) \end{aligned}$$

where p^c is the histogram bin index for a pixel p, and N is the total number of pixels for each image. 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

$$\begin{aligned} \pi_i &= \frac{\sum_j c_{ij}}{\sum_i \sum_j c_{ij}} \propto \sum_j c_{ij} & (12) \\ &\doteq \#(p^c = c_i) \times n_d \propto \#(p^c = c_i). & (13) \end{aligned}$$

It means that the MSF described in [4] 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 Fig. 1.

The success of MSF stems from its combination with the auto-correlogram features $\boldsymbol{\pi}^a$, and the weighted difference between these two types of features implicitly characterizes inter-bin spatial co-occurrence information.

3. DIRECTED MARKOV STATIONARY FEATURES

As proved above, the inherent ambiguity associated with MSF is caused by the symmetric property of the co-occurrence matrix. In this section, we present the so called Directed Markov Stationary Features (DMSF) for addressing this problem. Instead of using undirected pixel pairs, we extract the pixel pairs for each selected direction. Namely, we define several sets of local pixel pairs corresponding to different directions, i.e.,

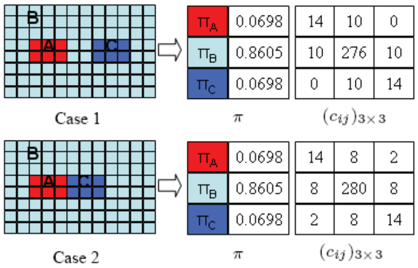

Fig. 1. An example which shows an informative trivial solution of MSF with no discriminant information. Note that we set $d = 1$ for this example.

horizontal $\{S_H\}$, vertical $\{S_V\}$ and two diagonal directions $\{S_{D1}\}$, $\{S_{D2}\}$ as illustrated in Fig. 2. We calculate the co-occurrence matrix for each selected direction. More specifically, the spatial co-occurrence matrix $\mathbf{C}^H$ modeled by all the horizontally directed pixel pairs is defined as

$$c_{ij}^H = \#(p_1^c = c_i, p_2^c = c_j \mid (p_1, p_2) \in \{S_H\}). \quad (14)$$

Similarly, for vertical or diagonal directed pairs, we calculate the co-occurrence matrix as:

$$c_{ij}^V = \#(p_1^c = c_i, p_2^c = c_j \mid (p_1, p_2) \in \{S_V\}). \quad (15)$$

$$c_{ij}^{D1} = \#(p_1^c = c_i, p_2^c = c_j \mid (p_1, p_2) \in \{S_{D1}\}). \quad (16)$$

$$c_{ij}^{D2} = \#(p_1^c = c_i, p_2^c = c_j \mid (p_1, p_2) \in \{S_{D2}\}). \quad (17)$$

Note that the matrices c_{ij}^H, c_{ij}^V, c_{ij}^{D1} and c_{ij}^{D2} are no longer symmetric since the pixel pairs are directed, namely, $c_{ij}^H \neq c_{ji}^H$. The definitions of p_{ij}^H, p_{ij}^V, p_{ij}^{D1} and p_{ij}^{D2}, are the same as in Eqn. (2). And we extract the stationary distribution $\boldsymbol{\pi}^H$, $\boldsymbol{\pi}^V$, $\boldsymbol{\pi}^{D1}$ and $\boldsymbol{\pi}^{D2}$ and the initial distribution $\boldsymbol{\pi}_a^H$,$\boldsymbol{\pi}_a^V$,$\boldsymbol{\pi}_a^{D1}$ and $\boldsymbol{\pi}_a^{D2}$ using the same method as in Eqn. (6), which gives us an $8K$ representation vector:

$$\mathbf{x} = [\boldsymbol{\pi}^H, \boldsymbol{\pi}^V, \boldsymbol{\pi}^{D1}, \boldsymbol{\pi}^{D2}, \boldsymbol{\pi}_a^H, \boldsymbol{\pi}_a^V, \boldsymbol{\pi}_a^{D1}, \boldsymbol{\pi}_a^{D2}], \quad (18)$$

where $\boldsymbol{\pi}_a^H$ is the auto-correlogram, namely the normalized version of the vector $[c_{11}^H, c_{22}^H, \cdots, c_{KK}^H]$. The significant difference with MSF is that now the stationary distribution $\boldsymbol{\pi}^H$, $\boldsymbol{\pi}^V$, $\boldsymbol{\pi}^{D1}$ and $\boldsymbol{\pi}^{D2}$ do not correspond to any informative trivial solutions, and it can truly characterize the spatial context information.

4. EXPERIMENTS

4.1. Data Sets

Two face databases CMU PIE [6] and FRGC V1.0 [7] are used in the face recognition experiments. We used 3329 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. Some sample face images from both databases are shown in Fig. 3.

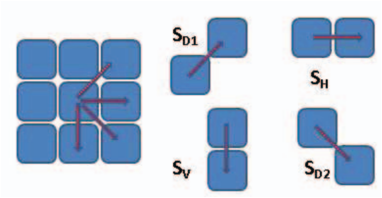

Fig. 2. Illustration of the way to extract different directed pixel pairs.

Fig. 3. Some sample face images from CMU PIE database (top row) and FRGC V1.0 database (bottom row)

4.2. Face Recognition based on DMSF

To validate the general performance of MSF and our proposed DMSF, we exploit three different low level features widely used for face recognition, namely, gray level, local binary pattern (LBP) [8], and direction of gradient (DOG) [9]. For gray level feature, 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. Note that the bin number we chosen is based on the best classification performance of each raw histogram representation respectively. We vary the size of the input image by down-sampling with bilinear interpolation to validate the robustness of the representations. For MSF, we reported both the approximation solution [4] and our derived row sum solution (i.e., MSF (V2)). We do not further compare the proposed DMSF with coherence vector and auto-correlogram features, since the MSF shows to be superior over them [4].

Note that although many stronger classification algorithms, *e.g.*, Support Vector Machine (SVM) [10], 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 affection of the consequent strong classifiers. The dissimilarity measurement is based on χ^2

Table 1. A summary of the recognition rates (%) for face classification on CMU PIE and FRGC V1.0 databases.

FRGC Ver-1.0 Dataset				CMU PIE Dataset			
Feature	Gray Level	LBP	DOG	Feature	Gray Level	LBP	DOG
Image Size: 100 × 100				Image Size: 64 × 64			
Histogram	34.50	39.45	27.29	Histogram	42.54	72.25	66.30
MSF	33.79	37.12	35.45	MSF	46.53	62.25	77.96
MSF (V2)	39.34	**40.37**	34.43	MSF (V2)	45.92	**74.52**	76.12
DMSF	**42.24**	36.69	**38.42**	DMSF	**50.95**	70.23	**82.81**
Image Size: 50 × 50				Image Size: 32 × 32			
Histogram	32.63	60.34	31.74	Histogram	37.45	76.55	59.98
MSF	34.29	57.65	40.76	MSF	40.21	72.87	67.77
MSF (V2)	36.87	62.57	39.98	MSF (V2)	41.50	77.04	67.22
DMSF	**45.49**	**63.06**	**48.25**	DMSF	**47.08**	**82.32**	**74.95**
Image Size: 25 × 25				Image Size: 16 × 16			
Histogram	24.67	54.44	21.63	Histogram	29.04	**66.30**	45.30
MSF	26.76	50.94	33.33	MSF	30.69	55.31	54.82
MSF (V2)	28.17	**58.43**	33.12	MSF (V2)	30.88	64.95	56.17
DMSF	**35.95**	55.04	**44.79**	DMSF	**37.02**	60.28	**66.97**

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

$$D(\mathbf{x}, \mathbf{y}) = \frac{1}{2} \sum_{j} \frac{(x_j - y_j)^2}{x_j + y_j}. \qquad (19)$$

The comparison results with the original histogram, MSF and DMSF are listed in Table 1, from which the following observations can be made: 1) The original histogram gives the lowest recognition rates, and MSF improves the performance for gray level and DOG features as reported in [4]; 2) Our proposed DMSF generally gives significant improvement on recognition accuracy compared with MSF, which validates the effectiveness of our method, since it avoids the inherent ambiguity of the trivial solution of MSF. 3) However, one could also see that for LBP features, sometimes the original MSF (i.e., polynomial approximation) and our proposed DMSF (i.e., which also adopts the approximation solution) gives poor performance. This is due to the fact that for some certain visual feature quantization, a considerable number of histogram bins are empty, which results in unstable calculation of the approximation solution. One could observe that our derived solution of MSF (V2), i.e., using the row sum method, does not suffer from this issue. 4) The performance of our proposed DMSF is robust in terms of different image scales and different visual features.

5. CONCLUSIONS

In this paper, we addressed the inherent ambiguity of the recently proposed MSF method for image spatial context modeling. And we proposed the so called Directed Markov Stationary Features to eliminate this ambiguity as well as to incorporate more spatial context information. Our experimental results well justified the effectiveness of the proposed DMSF.

Acknowledgment

This work was supported by NRF/IDM grant R-263-000-524-279, Singapore.

6. REFERENCES

[1] E. Hadjidemetriou, M. Grossberg, and B. Nayar, "Multiresolution histograms and their use for recognition," *TPAMI*, vol. 26, pp. 831–847, 2004.

[2] S. Birchfield and S. Rangarajan, "Spatiograms versus histograms for region-based tracking," in *CVPR*. IEEE, 2005, pp. 1158–1163.

[3] J. Huang, S. Kumar, M. Mitra, W. Zhu, and R. Zabih, "Spatial color indexing and applications," *IJCV*, vol. 35, pp. 245–268, 1999.

[4] J. Li, W. Wu, T. Wang, and Y. Zhang, "One step beyond histogram: Image representation using markov stationary features," in *CVPR*. IEEE, 2008.

[5] L. Breiman, "Probability," *Society for Industrial and Applied Mathematics*, 1992.

[6] T. Sim, S. Baker, and M. Bsat, "The cmu pose, illumination, and expression database," *TPAMI*, vol. 25, pp. 1615–1618, 2003.

[7] 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," in *CVPR*. IEEE, 2005, pp. 947–954.

[8] T. Ojala, M. Pietikainen, and T. Maenpaa, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *TPAMI*, vol. 24, pp. 971–987, 2002.

[9] N. Dalai and B. Triggs, "Histograms of oriented gradients for human detection," in *CVPR*. IEEE, 2005, pp. 886–893.

[10] D. Meyer, F. Leisch, and K. Hornik, "The support vector machine under test," *Neurocomputing*, vol. 55, pp. 169–186, 2003.