One Step Beyond Histograms: Image Representation using Markov Stationary Features

Jianguo Li, Weixin Wu, Tao Wang, Yimin Zhang
Intel China Research Center, Haidian, Beijing, 100190
{jianguo.li, weixin.wu, tao.wang, yimin.zhang}@intel.com

Abstract

This paper proposes a general framework called Markov stationary features (MSF) to extend histogram based features. The MSF characterizes the spatial co-occurrence of histogram patterns by Markov chain models, and finally yields a compact feature representation through Markov stationary analysis. Therefore, the MSF goes one step beyond histograms since it now involves spatial structure information of both within histogram bins and between histogram bins. Moreover, it still keeps simplicity, compactness, efficiency, and robustness. We demonstrate how the MSF is used to extend histogram based features like color histogram, edge histogram, local binary pattern histogram and histogram of oriented gradients. We evaluate the MSF extended histogram features on the task of TRECVID video concept detection. Results show that the proposed MSF extensions can achieve significant performance improvement over corresponding histogram features.

1. Introduction

Histograms are a widely used tool in computer vision and pattern recognition community to represent, analyze and characterize various visual inputs [24, 10, 11, 8]. As an example of their applicability, color histograms were initially applied in areas like image/video retrieval [24]. Following that work, histograms are not only used to develop various vision systems [7, 6, 8], but also extended to other kinds of visual inputs. For example, histograms have been used to represent processed images such as image edges, image gradients and so on. In details, edge histograms have been used for image/video retrieval [19, 16], while histograms of gradients (HoG) are used in human detection [6]. Besides, histograms are also used as basic descriptor in local binary patterns (LBP) [18], SIFT [14], and even the Bag-of-Feature framework [5, 8, 13]. The reason of importance is that histograms are fast to compute, space efficient, and robust to noise [10].

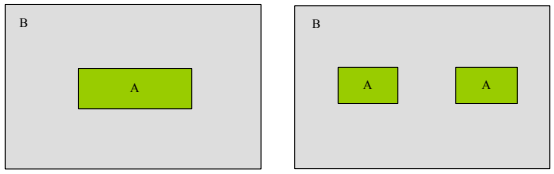


Figure 1. Two images have the same color histogram but different contents (A, B indicate two colors).

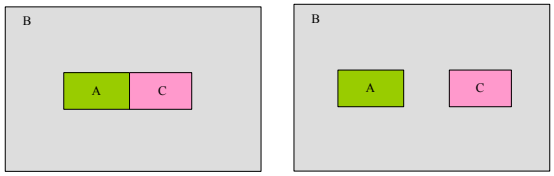
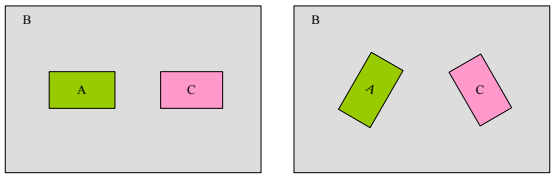


Figure 2. Images have the same intra-color structure but different extra-color structure (A, B, C indicate three colors).



(a) (b)

Figure 3. Histogram analysis undistinguishable examples.

Although histograms are widely applied, they are inadequate for many applications since they do not capture any spatial information. For example, color histograms cannot distinguish the two images in Figure 1. Further analysis on histogram bins will yield more powerful discrimination capability. In general, the discrimination capability of histogram analysis can be divided as follows:

I. Histogram-level distinguishable: images can be directly distinguished by their histogram representations;

II. Intra-bin distinguishable: images can be distinguished by spatial relationship analysis *within* each histogram bin. Figure 1 shows example images required this kind of distinguishable.

III. Extra-bin distinguishable: images can be distin-

978-1-4244-2243-2/08/$25.00 ©2008 IEEE

guished by spatial relationship analysis *between* histogram bins. Figure 2 illustrates example images required this kind of distinguishable.

IV. Histogram undistinguishable: images cannot be directly distinguished by (global) histogram analysis. Figure 3 illustrates such examples.

Most conventional histogram features can only handle level-I distinguishable problem, which greatly restrict their performance in vision systems and prevent them from being extensively used in real applications. Hence, different kinds of improvements were proposed to alleviate this limitation, especially for color histograms. Generally, the existing improvements can be grouped into two categories:

The first category focuses on directly adding spatial structure information to histograms. Color spatiogram adds lower order moments of spatial distribution to each histogram bin [1]. It has been successfully applied in object tracking. Coherence vectors divide pixels of each histogram bin into coherent pixels and non-coherent pixels so that they are able to distinguish different region connectedness within histogram bins [20]. Auto-correlogram describes spatial information by counting the spatial co-occurrence of histogram patterns [12]. Both coherence vectors and auto-correlogram achieve notable performance boost over histograms in real applications like image retrieval [12, 15]. However, all these extensions still belong to level-II extension since they still have difficulties in distinguishing images in Figure 2.

The second category includes extensions that aim to improve histogram-level distinguishing capability via pyramid or spatial layout. The pyramid/multiresolution histograms are proposed to represent image spatial information in a pyramid way [10, 11, 8, 13], while the layout extension extracts histograms from predefined layout grids. Their drawbacks are still obvious. The multiresolution histograms even have difficulties in distinguishing images in Figure 1. The layout extension only catches macro-level spatial information while still does not represent local structure well. Moreover, these two extensions are not as essential as those extensions in the first category, and can be applied to further extend the first-category extensions. Therefore, this paper focuses on essential extensions.

We propose a general framework called Markov stationary feature (MSF) that can essentially handle the three-level histogram distinguishable problems, and thus alleviate the limitation of histograms. The basic idea is that we adopt Markov chain models to characterize the spatial co-occurrence of histogram patterns, and further reduce the comparison between Markov chains to the corresponding initial distributions and stationary distributions. Hence, both intra-bin information and extra-bin information are compactly encoded. Our major contribution can be summarized as follows:

(1) Present a systematic categorization of histogram analysis, i.e., histogram-level distinguishable, intra-bin distinguishable, and extra-bin distinguishable.

(2) Propose the MSF framework that can handle the three-level distinguishable problems, while still keeps simplicity, compactness, efficiency, and robustness.

(3) Demonstrate how the MSF framework is used to extend histogram based representations such as color histogram, edge histogram, LBP histogram, and HoG.

The rest of this paper is organized as follows. Section 2 presents motivation, theoretic justification, and computing diagram of the proposed MSF framework. Section 3 presents how MSF is used to extend conventional histogram features. Section 4 discusses further possible enhancements. Section 5 demonstrates a state-of-the-art video concept detection system based on the MSF extended histogram features. Conclusions are drawn in the final section.

2. The Markov Stationary Feature Framework

Suppose visual inputs (raw or processed images) are quantized into K levels S=$\{c_1, \cdots, c_K\}$ (i.e., K histogram bins), this paper aims at a feature representation that can characterize both intra histogram-bin spatial information and extra histogram-bin spatial information.

2.1. Correlogram Revisited

Before diving into details, we'd like to revisit correlogram first, which is one of the motivations of our work [12].

The color correlogram of image I is defined as a squared table where the entry at $(i,\ j)$ specifies the probability of finding a pixel of color c_j at a fixed distance d from a given pixel of color c_i. Mathematically,

$$\begin{aligned}\gamma_{i,j}^{d}(I) &= Pr(p_2 = c_j \,|p_1 = c_i,\ |p_1 - p_2| = d) \\ &= \frac{\#(p_1 = c_i, p_2 = c_j \,||p_1 - p_2| = d)}{\#(p_1 = c_i)}, \end{aligned} \quad (1)$$

where $\#(p_1 = c_i)$ denotes the number of pixels falling into histogram bin c_i.

For K different colors, the full color correlogram contains K^2 elements that is not only space expensive but also sensitive to image noise. Hence, almost none real application directly uses the full correlogram. There is a widely used simplification that just uses the K diagonal elements of correlogram, namely color auto-correlogram (CAC). Although CAC demonstrates notable performance boost over color histogram in practice [12], it is still a level-II extension of histogram that cannot handle the extra-bin distinguishable problem illustrated in Figure 2.

It is interesting to notice that the correlogram is factually stemmed from the spatial co-occurrence matrix, which

contains rich information. The problem is that the full co-occurrence matrix is space expensive and sensitive to image noise. Is it possible to build a compact yet robust feature representation from the co-occurrence matrix? The MSF framework is the proposed answer for this question.

2.2. Motivations

Let p=(x, y) be a pixel in image I, the spatial co-occurrence matrix is defined as $C = (c_{ij})_{K\times K}$ where

$$c_{ij} = \#(p_1 = c_i, p_2 = c_j \,||p_1 - p_2| = d)/2, \qquad (2)$$

in which d indicates L_1 distance between two pixels p_1 and p_2, and c_{ij} counts the number of spatial co-occurrence for bin c_i and c_j [1]. Figure 4(a) illustrates an example for accumulating co-occurrence matrix C in a local area.

The spatial co-occurrence c_{ij} can be interpreted in a statistical view. When the pattern c_i and c_j have large spatial co-occurrence, then the possibility that c_i transits to c_j will be high. From this perspective, we adopt Markov chain models to characterize spatial relationship between histogram bins, which treats the bins as states in Markov chain models, and interprets the co-occurrence as the transition probability between bins. In this way, the comparison of two histograms are transferred to the comparison of two corresponding Markov chains. The following work will present theoretic background on how to compare two Markov chains in a robust and efficient way.

2.3. Theoretic Justification

A Markov chain [3] is a sequence of random observed variables with the Markov property, namely that, given the present state, the future and past states are independent. Formally, $P(X_{n+1}|X_n, \cdots, X_1) = P(X_{n+1}|X_n)$. All possible values of X_n form a countable set S called the state space of the chain. Suppose images are quantized into K levels, the state space can be denoted as $S = \{c_1, \cdots, c_K\}$.

In this paper, the state space is assumed fixed for all images. Therefore, a Markov chain will totally depend on two basic ingredients, namely a probability transition matrix P and an initial distribution $\pi(0)$. The transition probability going from state c_i to c_j is denoted as $p_{ij} = P(X_1 = c_i|X_2 = c_j)$, and the Markov transition matrix $P = (p_{ij})_{K\times K}$ (i-th row, j-th column) must obey the following two properties: (1)$p_{ij} \geq 0$, $\forall c_i \in S, c_j \in S$; (2) $\sum_{j=1}^{K} p_{ij} = 1$.

According to our interpretation and the properties of the Markov transition matrix, it can be constructed from the spatial co-occurrence matrix $C = (c_{ij})_{K\times K}$ by

$$p_{ij} = c_{ij} / \sum_{j=1}^{K} c_{ij}. \qquad (3)$$

[1]Note that $d = 1$ is used in all our illustrations and experiments, and the division of 2 means that we do not repeatedly accumulate c_{ij} and c_{ji}.

The direct thought is comparing two Markov chains by comparing their transition matrices. However, the transition matrix is sensitive to image noise and space expensive (requires $O(K^2)$ space for K states). It is expected to achieve a space efficient and robust solution based on properties of Markov chains.

Suppose the state distribution after n steps is $\pi(n)$ and the initial distribution is $\pi(0)$ (row vectors), the Markov transition matrix obeys the following rules (Here only lists theoretic results, please refer to [3] for more details):

Lemma 1: (1)$\pi(n+1) = \pi(n)P$, $\pi(n) = \pi(0)P^n$; (2)$P^{m+n} = P^n P^m$, where P^n is the n-th order power of the transition matrix P.

The equation in the second property is known as the *Chapman-Kolmogorov equation*. With the state transition rule in Lemma 1, we have a very useful definition:

Definition 1: A distribution π is called *a stationary distribution* when it satisfies $\pi = \pi P$.

According to the Chapman-Kolmogorov equation, it is obvious that for a stationary distribution, $\pi = \pi P = \cdots = \pi P^n$. Hence, the stationary distribution is known as an invariant measure of a Markov chain. Our intuitive idea is to adopt the stationary distribution as the compact representation of the Markov chain. However, we must guarantee the existence and uniqueness of the stationary distribution for any Markov chains. Concretely, the problem should be analyzed for two different cases.

2.3.1 Case 1: Regular Markov Chains

Before going deep into details, we present some basic definitions in Markov chains.

Definition 2: (1) A Markov chain is said to be *irreducible* if every state is accessible from every other state. (2) A process is *periodic* if there exists at least one state to which the process will continually return with a fixed step period (greater than one). *Aperiodic* means that there is no such state. (3) A chain is called *positive recurrent* when it can return each state within finite steps in average.

As to the existence and uniqueness of the stationary distribution of Markov chains, we have the following lemma.

Lemma 2: (1) If the Markov chain is *positive recurrent*, there **exists** a stationary distribution. (2) If the chain is *positive recurrent* and *irreducible*, there exists a **unique** stationary distribution.

Definition 3: A *positive recurrent* and *irreducible* Markov chain is called a *regular Markov chain*.

The unique stationary distribution of a regular Markov chain can be computed via the following theorem [3].

Theorem 1: For a regular Markov chain, we have $\lim_{n\to\infty} P^n = W$, where each row of matrix W is the same strictly positive probability vector $\vec{w}$ (i.e., the elements are all positive and sum to 1). Furthermore, $\vec{w}$ obeys $\vec{w}P = \vec{w}$.

It is obviously that $\vec{w}$ in this theorem is the stationary distribution of the corresponding regular Markov chain.

As an example, given a quantized color image, the quantized color space (or state space) is obvious irreducible since each pixel is connected by neighbor pixels, and there are no isolated pixels. However, we cannot guarantee the state space is positive recurrent, i.e., we cannot guarantee the Markov chain is regular.

2.3.2 Case 2: General Markov Chains

For an irregular Markov chain, the n-th step transition matrix P^n may be periodic (i.e., after every fixed m steps, $P^{n+m} = P^n$). To remove periodicity in computing stationary distribution, we may resort to the fundamental limitation theorem of Markov chains [3].

Theorem 2: (1) The limitation $A = \lim_{n\to\infty} A_n$ **exists** for all state-countable Markov chains, where $A_n = \frac{1}{n+1}\big(I + P + P^2 + \cdots + P^n\big)$, and I is an identity matrix. (2) When the chain is regular, each row of matrix A is equal to the unique stationary distribution vector.

In brief, it is consistent to compute the unique stationary distribution for any (regular and irregular) Markov chains via Theorem 2. In practice, we first choose a proper n (i.e., n=50), and approximate matrix A by $A_n = \frac{1}{n+1}\big(I + P + \cdots + P^n\big)$. In order to further reduce approximation error when using small n, according to the fact that each row of A should be the same, the stationary distribution is approximated by the average of each row $\vec{a}_i$ of A_n:

$$\pi \approx \frac{1}{K}\sum_{i=1}^{K} \vec{a}_i, \quad \text{where } A_n = [\vec{a}_1, \cdots, \vec{a}_K]^T. \qquad (4)$$

2.4. Feature Formulation

Previous subsection presents theoretic justification that stationary distribution is a unique and invariant measure of the Markov transition matrix. In practice, the Markov stationary feature is defined as the combination of the initial distribution $\pi(0)$ and the stationary distribution π. The initial distribution cannot be ignored since a Markov chain is determined by both its transition matrix and by its initial distribution. More detailed, we cannot guarantee that initial distributions are the same for all images, and cannot guarantee that the Markov chain is a strict stationary process for all images (i.e., $\pi(0) = \pi$ for all images) . The combination can also be interpreted from the following perspectives:

(1) The initial distribution encodes the intra-bin transitions (self-transitions).

(2) The stationary distribution further encodes the extra-bin transitions.

We can see that the MSF framework encodes both intra-bin and extra-bin structure information. This is a great advance over conventional histogram features and existed extensions. Besides, the feature is space efficient ($2\times K$ in feature dimension), that is comparable to histogram, but has significantly lower space requirement than full transition matrix (K^2 in the same case). Moreover, since MSF comes from the stationary distribution, it will be more robust to image noise than full transition matrix.

2.5. Computing Diagram and Examples

According to the feature formulation, the MSF feature can be computed straightforward in 5 steps as follows.

Table 1. The computing scheme for the MSF feature

S1:	Quantize visual inputs into K levels such as color space quantization for color histogram;
S2:	Given a defined distance d, accumulate the spatial co-occurrence matrix $C = (c_{ij})_{K\times K}$ by Equation (2);
S3:	Calculate the Markov transition matrix $P = (p_{ij})_{K\times K}$ from $C = (c_{ij})_{K\times K}$ by Equation (3);
S4:	Compute the stationary distribution π according to Equation (4);
S5:	Normalize the self-transition as the initial distribution $\pi(0)$, and combine it with the stationary distribution π to obtain the complete MSF feature $\vec{h}_{MSF} = [\pi(0), \pi]^T$.

To explicitly demonstrate the advantages of the MSF feature, Figure 4 illustrates how the MSF-color feature is computed, and how the MSF-color feature beats traditional color descriptor in distinguishing images in Figure 2.

Suppose the histogram values of three colors in the two images are h_A, h_B and h_C, the color spatial co-occurrence matrices are illustrated in the first column of Figure 4(b). According to the definition of color auto-correlogram (CAC), it is not hard to find that two images in Figure 2 have the same CAC feature, i.e., $[160/h_A, 340/h_B, 160/h_C]^T$. Note that the initial distributions of CAC and MSF are slightly different since they are normalized by different denominators. The middle-column of Figure 4(b) shows the Markov transition matrix of the two images. Note that the transition matrix may not be symmetrical. Figure 5 shows the graph of one-step Markov chain according to the transition matrix. The stationary distributions are computed according to Equation (4) and illustrated in the right most column of Figure 4(b). It is obvious that these two images have

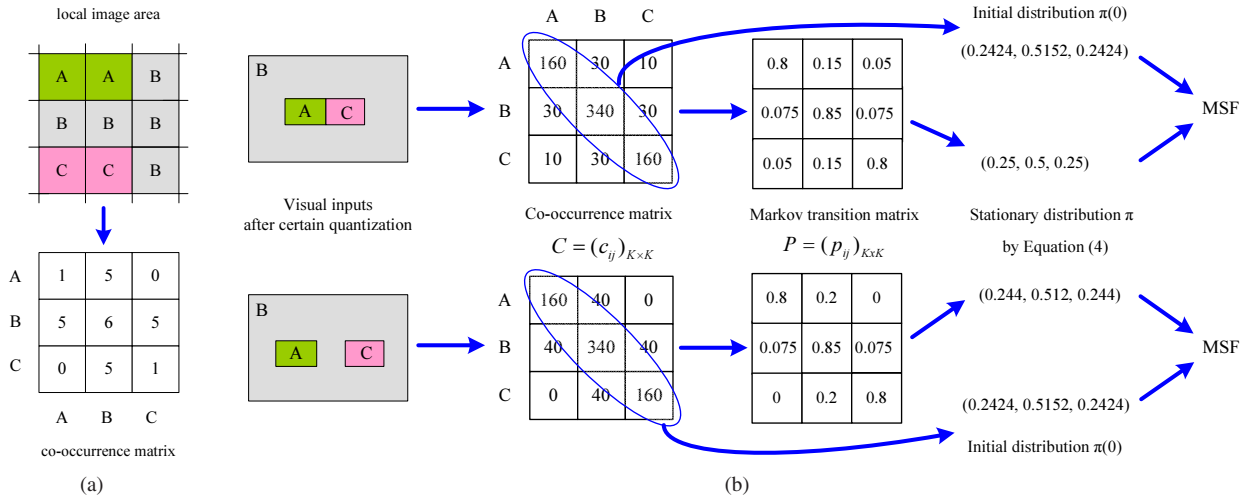


Figure 4. (a) Co-occurrence matrix in a local area; (b) How MSF-Color is computed for images in Figure 2 (A, B, C indicate 3 colors).

different stationary distribution. This case clearly shows images that the MSF extended color descriptor can distinguish while color histogram and CAC cannot.

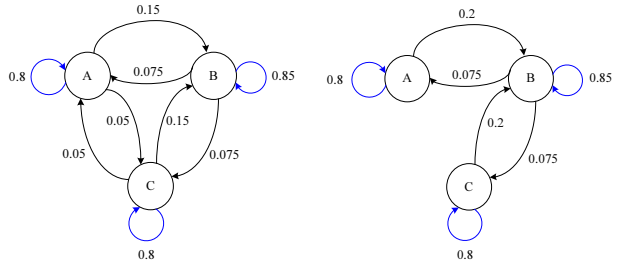


Figure 5. One step Markov chains for images in Figure 2, which characterize the color spatial co-occurrence (A, B, C indicate three colors in the images or states of corresponding Markov chains).

2.6. Similarity Measures

Distance/similarity measures can be defined over the MSF feature to realize image comparison, matching, retrieval, etc. As elements in MSF feature keep non-negative, the distance metrics used in histograms can be borrowed. For example, histogram intersection distance, χ^2 distance, Euclidean distance, etc [27]. In this paper, we adopt the well known χ^2 distance as suggested by [12, 27].

Suppose there are two images A and B, their MSF initial distributions are $\vec{h}_I(A)$ and $\vec{h}_I(B)$ respectively, and their stationary distributions are $\vec{h}_S(A)$ and $\vec{h}_S(B)$ respectively. The MSF features of the two images are $\vec{h}(A) = [\vec{h}_I(A), \vec{h}_S(A)]^T$, and $\vec{h}(B) = [\vec{h}_I(B), \vec{h}_S(B)]^T$. The χ^2 distance between $\vec{h}(A)$ and $\vec{h}(B)$ is defined as

$$D(\vec{h}_A, \vec{h}_B) = D(\vec{h}_I(A), \vec{h}_I(B)) + D(\vec{h}_S(A), \vec{h}_S(B)), \quad (5)$$

$$D(\vec{h}_I(A), \vec{h}_I(B)) = \sum_{j=1}^{K} \frac{\left[h_{Ij}(A) - h_{Ij}(B)\right]^2}{h_{Ij}(A) + h_{Ij}(B)},$$

where $\vec{h}_I(A) = [h_{I1}(A), h_{I2}(A) \cdots, h_{IK}(A)]^T$, $\vec{h}_S(A) = [h_{S1}(A), h_{S2}(A) \cdots, h_{SK}(A)]^T$.

This distance can be easily embedded into the RBF kernel when using kernel based methods such as support vector machines [25]: $k(\vec{h}_A, \vec{h}_B) = \exp\{-\gamma \cdot D(\vec{h}_A, \vec{h}_B)\}$, where γ is the parameter of RBF kernel.

3. MSF Extended Histogram Features

The MSF presents a general way for extending histogram based features. In this paper, we employed the MSF framework to extend four kinds of histogram features: color histogram, edge histogram, LBP histogram, and gradient histogram (a.k.a. HoG). As the MSF can be computed straightforward via the scheme in Table 1, here we only need to show how the co-occurrence matrix is defined for various visual inputs.

MSF-Color: The MSF extension of color histogram is called MSF-Color. To build MSF-Color, input image should be transformed to a proper color space and quantized according to a suitable codebook [24]. As suggested by [22], the 166-level quantized HSV (hue, saturation, value) color space was used in our practice for this purpose. And then the co-occurrence matrix is computed according to Equation (2).

MSF-Edge: The MSF extension of edge histogram is called MSF-Edge. Given raw image I, the edge image E was extracted by Canny edge detector, in which $E(x, y)>0$ if pixel (x, y) is on an edge, otherwise 0. It is obvious that edge image E is a sparse image. This paper quantized the edge image into K=64

level according to edge pixel orientation and magnitude (8 orientations × 8 magnitudes). Given the bin set $S = \{s_1, \cdots, s_K\}$, the edge co-occurrence matrix is defined as $C_E = (e_{ij})_{K\times K}$, where $e_{ij} = \#\big(E(p_1) = s_i, E(p_2) = s_j \big| |p_1 - p_2| \leq d\big)$, in which $E(p_1)$ indicates quantized value at edge pixel p_1. Note that in practice, we performed MSF-Edge feature extraction on an widely used layout (2×2 grids plus an overlapped center grid) for better performance.

MSF-HoG: The MSF extension of HoG is called MSF-HoG. Suppose G is the gradient image of raw image I, i.e., $G(x, y)$=$(\partial I/\partial x, \partial I/\partial y)$, the HoG quantizes the gradient image according to pixel gradient orientations into 64 levels [6]. As G is a full image, MSF-HoG follows the same computing scheme as MSF-Color. Note that in practice, we also extracted MSF-HoG feature using the same grid layout as in MSF-Edge.

MSF-LBP: LBP operator is a theoretically simple yet powerful method for analyzing textures. It will yield a 256-level quantized image when manipulating LBP operator on the raw input image. Enhanced LBP operators lead to lower levels. For example, uniform LBP operator yields 59 levels [18]. The MSF extension of LBP is called MSF-LBP. As the LBP operation generates a full image, MSF-LBP follows the same computing scheme as MSF-Color. Note that in practice, we performed uniform LBP operation in HSV color space, and concatenated results from three color channels together for the final representation.

4. Discussions

We have shown some image examples in Figure 3 which are histogram analysis undistinguishable. In fact, histograms are low sensitive or invariant to some certain types of transformations. For example, rotation, translation of objects; permutation of image pixels; and so on. Ref.[9] studies the histogram preserving image transformations. It would be interesting to further study which kinds of transformations are MSF preserving.

If it is required to alleviate the impact of some MSF preserving transformations, the MSF feature can be enhanced by spatial layout extensions or pyramid/multi-resolution extensions. The layout extension extracts the MSF feature from some given grid layout of images or segmented regions, while the multi-resolution extension can be done in two different ways. One is constructing multi-resolution co-occurrence matrix by a series of contiguous banded neighborhood definition as in [12]: $C^{d_1,d_2}(I) = (c'_{ij})_{K\times K}$, where $c'_{ij} = \#(p_1 = c_i, p_2 = c_j \mid d_1 \leq |p_1 - p_2| \leq d_2)$. The other is pyramid filtering the raw input images to get a series of blurred images [10], and apply MSF feature extraction on each blurred image. Furthermore, it is possible to combine pyramid and spatial layout together, which yields a spatial pyramid representation [13, 2].

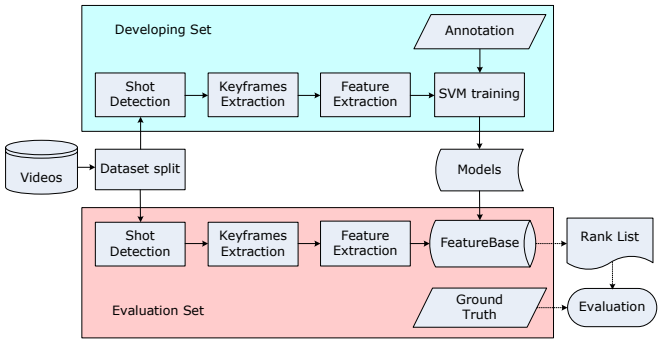


Figure 6. Evaluation flowchart of video concept detection system.

5. Applications for Video Concept Detection

This section will demonstrate a state-of-the-art video concept detection system based on the MSF extended histogram features. We first illustrate the evaluation system flowchart, and then show compared performance by both the conventional and the MSF extended histogram features.

5.1. Video Concept Detection in TRECVID

The video concept detection focuses on automatic video shots annotation by predefined concept lexicon, i.e., whether certain concept is presented in a given video shot or not. Our video concept detection system is evaluated on TRECVID (TREC video retrieval evaluation) data sets, which contain not only hundreds of hours of TV news or documentary video corpus, but also well defined concept lexicon. Moreover, there is annual TRECVID evaluation held by NIST (National Institute of Standards and Technology), who provides a systematic protocol for evaluating video concept detection performance [17, 21], and also for promoting related researches in computer vision, multimedia and machine learning. This evaluation is well received, and attracts participants from not only university research groups worldwide but also many industry labs [21].

Figure 6 illustrates the evaluation flowchart of video concept detection. The video corpus was first divided into developing set and evaluation set. Then shot boundaries were detected for all videos, and visual features were extracted from keyframes in each video shot. A binary SVM classifier was trained for each concept from the annotated samples. RBF kernels were adopted for all concepts, and the kernel parameters were well tuned by cross-validation. Furthermore, detectors were trained with probabilistic output [4] so that they can output rank list for the standard evaluation. As this study focuses more on low-level features, please refer to our TRECVID report for more details on detectors [26]. The final detection performance was measured by the average precision (AP) of the top 2,000 retrieved shots (standard

evaluation metric used in TRECVID 05 and before). In case that the evaluation set was not fully annotated, the inferred AP was used as the performance metric (standard metric used in TRECVID 2007) [21].

5.2. Experimental Results

We evaluate the MSF extended histogram features on the TRECVID video concept detection task in both the 2005 video corpus and the 2007 video corpus. Generally, the 2005 video corpus consists of about 170 hours of TV news videos from different programs in English, Chinese and Arabic, while the 2007 video corpus consists of more than 100 hours of documentary videos in Dutch.

For the 2005 corpus, we adopted only the developing set as our target database since it contains full annotation for all 39 concepts such that we can obtain a comprehensive insight of the detection performance. Therefore, the 2005 developing corpus was further divided into two parts: the video sequences 141∼240 as the training set, and the video sequences 241∼277 as the testing set. This partition is the same as the MediaMill challenge [23], and thus we can compare our results directly to the MediaMill baseline. More detailed, the partition yields that the training set has 31,594 video shots, while the testing set contains about 12,313 video shots. The whole evaluation followed the flowchart illustrated in Figure 6. In this evaluation, we conducted experiments to compare the MSF extended histogram features to the corresponding conventional ones. To make a fair comparison, the original and extended features adopted the same quantization method and the same setup. Specially, we also did comparison between MSF-Color and other known extensions for color histogram, such as color coherence vectors (CCV) and color auto-correlogram (CAC). The averaged performance on all 39 concepts are listed in Figure 8(a). From the results, we can see that the performance gain ratio ranges from 5% to 35%, which is quite significant. Specially, the best extension (MSF-LBP) achieves 25% performance improvement over the MediaMill's baseline, which is in fact fusion results by several low level features such as Gabor, SIFT and so on. As the space is limited, we do not list results for each concept. But as an example, detail results on the concept "Car" is illustrated in Figure 7.

For the 2007 corpus, we just followed the standard partition for training and testing, in which there are 18,120 shots in the developing set, and 17,986 shots in the evaluation set. Note that the 2007 corpus is quite different from the 2005 corpus in two aspects. First, the video domain is changed from TV news to documentary in 2007, which yields quite different concept occurrence distribution in these two years. Second, the training set is much smaller than 2005. Both points make the 2007 corpus much harder than that of 2005. The final results also confirm this point. Nevertheless, the proposed MSF extensions still dominate the corresponding conventional histogram features. Figure 8(b) illustrates the averaged performance on all the 20 evaluated concepts. The performance gain ratio is ranging from 10% to 170%. In fact, the best extension here achieves comparable results with those top results in TRECVID 2007 evaluation, which were all obtained by fusing over many low-level features (some by more than 20 low-level features).

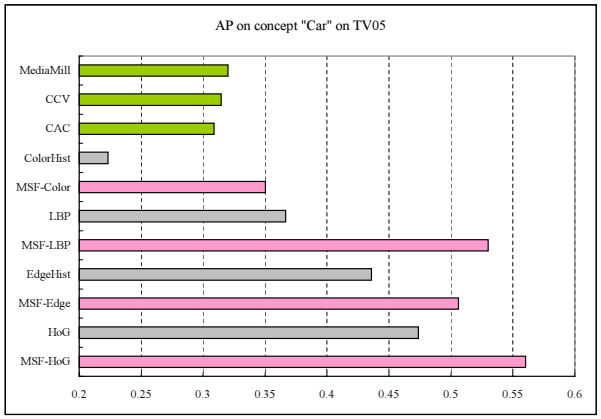


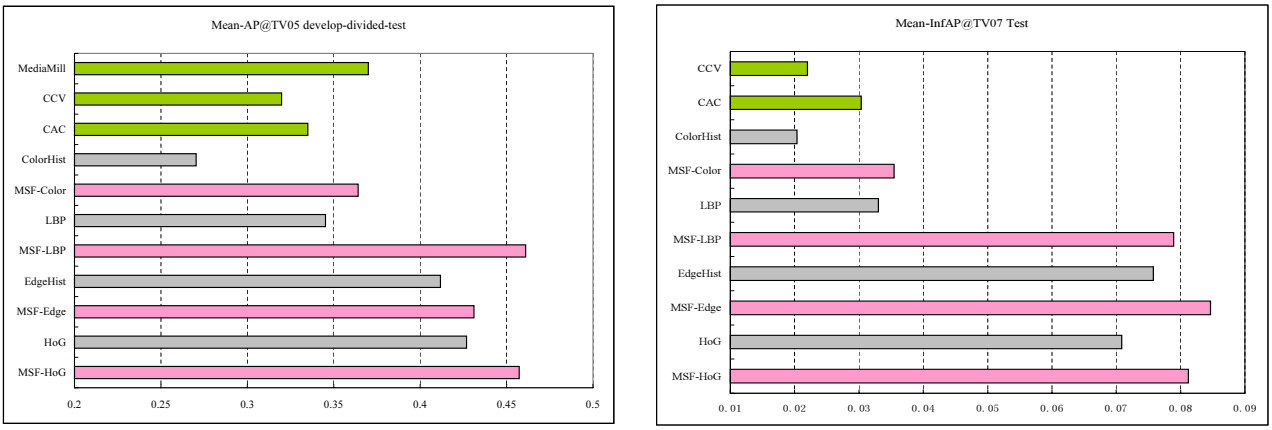
Figure 7. The performance on the concept "Car" by each compared feature on TRECVID 2005 corpus.

6. Conclusions

In this paper, we propose a general framework called Markov stationary features (MSF) to extend widely used histogram features. The MSF goes one step beyond histograms since it incorporate spatial structure information, and is able to handle all three levels of histogram analysis distinguishable problems. Moreover, it still keeps compactness, efficiency, and robustness. We present how the MSF framework is used to extend histogram based features such as color histogram, edge histogram, HoG and LBP histogram. We further demonstrate a state-of-the-art video concept detection system based on the MSF extended histogram features. Experiments on TRECVID data sets show that the MSF extensions can achieve significant performance improvement over corresponding conventional histogram features.

Our future work will consider the following two possibilities. First, we will try to employ MSF extended feature to other applications. For example, it is promising to try MSF-HoG in object detection for possible performance boost. Second, we will continue extending the MSF framework to other histogram based representation to incorporate spatial structure information. For instance, it would be interesting to employ the MSF philosophy to the bag-of-feature framework.

Acknowledgements: The authors wish to thank our U.S. colleagues Wu Yi and Haussecker Horst, and Yan



(a) (b)

Figure 8. The average performance of all concept detection results by each compared feature. (a) average performance on TRECVID 2005 corpus, (b) average performance on TRECVID 2007 corpus.

Shuicheng from National University of Singapore for helpful discussions, as well as the intelligent multimedia group of Tsinghua university for collaboration on TRECVID evaluation. Thanks also go to anonymous reviewers for their valuable suggestions and comments.

References

[1] S. Birchfield and S. Rangarajan. Spatiograms versus histograms for region-based tracking. In *CVPR*, pages 1158–1163, 2005. 2

[2] A. Bosch, A. Zisserman, and X. Munoz. Representing shape with a spatial pyramid kernel. In *Proc. ACM CIVR*, pages 401–408, 2007. 6

[3] L. Breiman. *Probability*. reprinted by SIAM, 1992. Chapter 7. 3, 4

[4] C.-C. Chang and C.-J. Lin. LIBSVM: a library for support vector machines, 2001. 6

[5] G. Csurka, C. Bray, and et al. Visual categorization with bags of keypoints. In *ECCV workshop on Statistical Learning in Computer Vision*, 2004. 1

[6] N. Dalai and B. Triggs. Histograms of oriented gradients for human detection. In *CVPR*, pages 886– 893, 2005. 1, 6

[7] M. Flickner and et al. Query by image and video content: the QBIC system. *IEEE Computer*, 28(9):23–32, 1995. 1

[8] K. Grauman and T. Darrell. The pyramid match kernel: Discriminative classification with sets of image features. In *ICCV*, pages 1458–1465, 2005. 1, 2

[9] E. Hadjidemetriou, M. Grossberg, and S. Nayar. Histogram preserving image transformations. *IJCV*, 45:5–23, 2001. 6

[10] E. Hadjidemetriou, M. Grossberg, and S. Nayar. Multiresolution histograms and their use for recognition. *IEEE PAMI*, 26(7):831–847, 2004. 1, 2, 6

[11] D. Heeger and J. Bergen. Pyramid-based texture analysis/synthesis. In *SIGGRAPH*, pages 229–238, 1995. 1, 2

[12] J. Huang, S. Kumar, and et al. Spatial color indexing and applications. *IJCV*, 35(3):245–268, 1999. 2, 5, 6

[13] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: spatial pyramid matching for recognizing natural scene categories. In *CVPR*, pages 2169–2178, 2006. 1, 2, 6

[14] D. Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 60(2):91–110, 2004. 1

[15] W.-Y. Ma and H. Zhang. Benchmarking of image features for content-based retrieval. In *IEEE Conf. Signals, Systems & Computers 1998*, pages 253–257, 1998. 2

[16] B. Manjunath, J.-R. Ohm, and et al. Color and texture descriptors. *IEEE CSVT*, 11(6):703–715, 2001. 1

[17] NIST. TREC video retrieval evaluation. http://www-nlpir.nist.gov/projects/trecvid/. 6

[18] T. Ojala, M. Pietikainen, and et al. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE PAMI*, 24(7):971–987, 2002. 1, 6

[19] D. Park, Y. Jeon, and C. Won. Efficient use of local edge histogram descriptor. In *Proc. ACM Multimedia*, pages 51–54, 2000. 1

[20] G. Pass, R. Zabih, and J. Miller. Comparing images using color coherence vectors. In *Proc. ACM Multimedia*, pages 65–73, 1997. 2

[21] A. Smeaton, P. Over, and W. Kraaij. Evaluation campaigns and TRECVID. In *ACM workshop on Multimedia Information Retrieval*, pages 321–330, 2006. 6, 7

[22] J. Smith and S. Chang. Visually searching the web for content. *IEEE Multimedia*, 4(3):12–20, 1997. 5

[23] C. G. Snoek, M. Worring, and et al. The challenge problem for automated detection of 101 semantic concepts in multimedia. In *Proc. ACM Multimedia*, pages 421–430, 2006. 7

[24] M. Swain and B. Ballard. Color indexing. *IJCV*, 7(1):11–32, 1991. 1, 5

[25] V. N. Vapnik. *The Nature of Statistical Learning Theory*. Springer press, 1995. 5

[26] J. Yuan and et al. THU and ICRC at TRECVID 2007. In *Proc. TRECVID workshop*, 2007. 6

[27] D. Zhang and G. Lu. Evaluation of similarity measurement for image retrieval. In *Int'l. Conf. Neural Networks and Signal Processing*, pages 928–931, 2003. 5