A Visual Attention Model for Video Based on Non-Negative Matrix Factorization Sparseness on Parts

Jianlong Zhang Xinbo Gao Song Xiao and Ming Tong
Xidian University, Xi'an Shaanxi Province,710071
jlzhang@mail.xidian.edu.cn Tel:86-29-88201956
xiaosong@mail.xidian.edu.cn Tel:86-29-88201956

Abstract—**Visual attention is one of the most important mechanism of HVS (human visual system) and has been applied into many fields. Research on visual attention model is hot and difficult. This paper presents a novel visual attention model for video based on NMFSCP (non-negative matrix factorization sparseness on parts). Saliency map of this model is generated by utilizing four types of visual attention features such as intensity, color, orientation and motion. Motion feature of video key frame is extracted by the NMFCP (Non-negative matrix factorization sparseness on parts) algorithm. Intensity, color and orientation features were obtained by the Itti visual attention model. Four features are combined with unequal linear coefficients according to the ratio of motion block in video frame. Simulation result shows the efficiency of proposed model.**

I. INTRODUCTION

With the rapid development of science and technology, intelligent processing is the important direction in application of military, industry and consumer etc. HVS (Human Visual System) plays an important role in our life and is researched to imitate the mind mechanism to realize intelligent system by establishing the mathematics model. As one of main contents of HVS, visual attention is taken more and more concentration in application of image compression, image retrieval, video analysis etc.

There are two types of Visual attention computation model, first one is Bottom-Up model based data driven which is low-level visual perception, second is Top-Down model which is high-level visual perception. At the beginning, visual attention model aimed at the static image, in visual attention model saliency map is used to interpret the saliency of visual region. Saliency map was proposed firstly by Koch and Ullman at 1985[1]. Their model fused the image features of color, intensity and orientation into saliency map, taking advantage of mechanism with inhabitation of return and WTA(Winner Take All).1998 Itti[2] proposed a classical Bottom-Up model based the model of Koch and Ullman adapted to nature image computation. Researchers proposed many models for static image and video processing based on Itti model. Zhang Longfei[3] presented a novel computable visual attention model for video skimming algorithm, this model adopted the alive-time of a visual objects as a new descriptor to improve the accuracy of locating highlight in a video clip, it used attention window(AW) and attention values of visual objects(AOs) to generate the attention curve of video. The algorithm can short 15%~25% time than traditional way. Junyong You[4] proposed to evaluate video quality by balancing two quality components: global quality and local quality using image quality metrics(IQM) with average spatiotemporal pooling, Saliency ,motion and contrast information are taken into account in modeling visual attention. The video quality modeling algorithm can improved the performance of image quality metrics on video quality assessment compared to the normal average spatiotemporal pooling scheme. Zhang Hua[5] proposed a distortions-weighing spatiotemporal visual attention model for the purpose of extracting the attention regions from distorted videos. Experimental result show that the model can not only accurately analyze the spatiotemporal saliency based on the intensity, texture and motion features but also able to estimate the blockiness of distortions in comparing with Walther's and You's models [6][7] .

Although all of the models for video mentioned above can get the saliency map of video frame, but they didn't consider accuracy of motion information extracted and real time capability of system processing. So we proposed a novel visual attention model based on NMFSCP algorithm for video. NMFSCP algorithm is described to extract the motion feature of video key frame in Section II and III. We will give the dataflow and structure of visual attention model of this paper in section IV. Simulation result will be section V. We will talk about the feature work in section VI.

II. NON-NEGATIVE MATRIX FACTORIZATION BASED ON SPARSENESS CONSTRAINT

A. *Introduction of non-negative matrix factorization(NMF)*

NMF has become increasingly popular for feature extraction in machine learning, computer vision, and signal processing. One reason for this popularity is that NMF codes naturally favor sparse, parts-based representation which in the context of recognition can be more robust that non-sparse, global features. NMF can factorize the signal into the weighted linear sum of group of base signals, and video frames can be considered the weighted linear sum of static part and motion part. In common, static part is non-sparse and motion part is sparse, so motion part can be extracted with NMF method. Here, we present a method of NMF sparseness constraint on parts, namely making sparseness constraint on part base vector of base matrix. This method can extract the motion information of video effectively.

B. Model of NMFSCP

Sparseness matrix denotes that most elements of matrix are zero or near zero, we use a sparseness measure based on the relationship between the $L1$ norm and the $L2$ norm:

$$spareness\left(y\right)=\frac{\sqrt{n}-\left(\Sigma\left|y_i\right|\right)/\sqrt{\Sigma y_i^2}}{\sqrt{n}-1} \quad (1)$$

Where n is the dimensionality of vector y [8]. This function evaluates to unity if and only if y contains only a single non-zero component, and takes a value of zero if and only if all components are equal(up to signs), interpolating smoothly between the two extremes.

Algorithm of NMFSCP can be described as optimization problem with constraint: given non-negative matrix $B \in R_+^{a\times b}$, solute a matrix of basis functions $W \in R_+^{m\times r}$ and coefficient matrix $H \in R_+^{r\times b}$, where r is the size of basis vector. $\| B \|_F$ is the Frobenius norm for matrix $B : \| B \|_F^2 = \sum_{ij} B_{ij}^2$ .So we define the NMFSCP problem as follows:

$$\begin{aligned} &\min_{W,H} \| B-\left(WH\right) \|_F^2 \\ &s.t. \quad W \geq 0, H \geq 0 \\ &\quad spareness\left(w_i\right)=s_i,\ \ i=1,2,\cdots z(z \leq r) \end{aligned} \quad (2)$$

where x_i is the i^{th} column of matrix W , S_i is size of sparseness expected, z is the number of vector which is added by sparseness constraint.

Rule iteration of Basis matrix W can be described as algorithm:

1. Initialize W and H to random positive matrices

2. Project each column of W to be non-negative, have unchanged $L2$ norm, but $L1$ norm set to achieve desired sparseness.

$L\left(x_i\right)$;

where $L\left(x_i\right)$ denote operator of nonlinear projection[8].

3. Iterate

i. Set $W := W - \lambda_w(WH-B)H^T$

ii. Make the each element of W to be non-negative,

$$W_{ij}=\begin{cases} W_{ij}, if \quad W_{ij} \geq 0 \\ 0 \quad , else \end{cases}$$

iii. Project each column of W to be non-negative , have unchanged $L2$ norm, but $L1$ norm set to achieve desired sparseness.

Rule iteration of coefficient matrix H can be described as formula (5):

$$H_{ij}=\frac{H_{ij}\sum_k\left(X_{ki}B_{kj}\right)}{\sum_k\left[X_{ki}\left(XH\right)_{kj}\right]} \quad (5)$$

III. Extraction of Video Motion Feature

Supposed that there are V frames with $M \times N$ resolution and the motion of video is continuous in video, we make sure one destination frame to be extracted every L frames, so we regard all between two destination frame as group of video(GOV). Then each frame of GOV is expanded by column as a column of matrix B , column size of B is $M \times N$ and row size is L , so three-dimensional video is transformed into two-dimensional non-negative Matrix. Then we use the algorithm of NMFSCP to factorize the matrix B, assume the factorization dimension of B is r and we just only give any of $r-1$ base vectors to add the sparseness constraint, so base vector $x_i\left(i=1,2,\cdots,r-1\right)$ with sparseness constraint added to factorization result represent the motion feature of video [8]. We can get motion feature of destination frame use formula (6):

$$M=\sum_{i=1}^{r-1} x_i H_{i,l+1} \quad (6)$$

Here, $H_{i,l+1}$ is weighted coefficient of key frame corresponding base vector by x_i .

Fig.1 (a), (b), (c), show the original image and extracted motion feature of 16th, 26th and 36th frame in video using the NMFSCP. We can draw a conclusion that NMFSCP can extract the motion feature of video accurately.

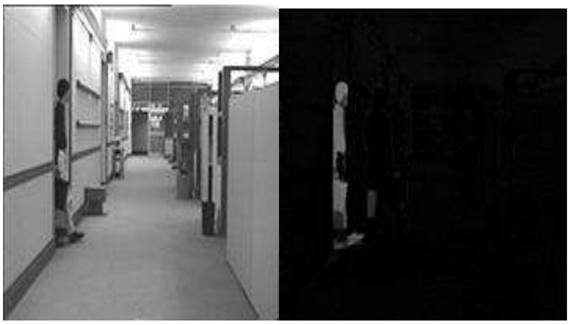

Fig.1 (a) original image and motion feature of 16th frame

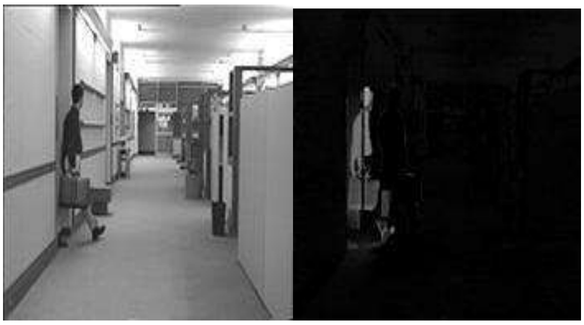

Fig.1 (b) original image and motion feature of 26th frame

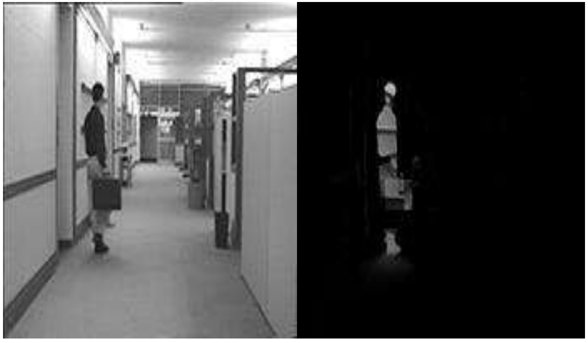

Fig.1 (c) original image and motion feature of 36thframe

IV. VISUAL ATTENTION MODEL PROPOSED THIS PAPER

This paper presents a visual attention model for video based NMFSCP algorithm and Itti static visual attention model. Fig.2 shows the structure and dataflow of attention model proposed by this paper.

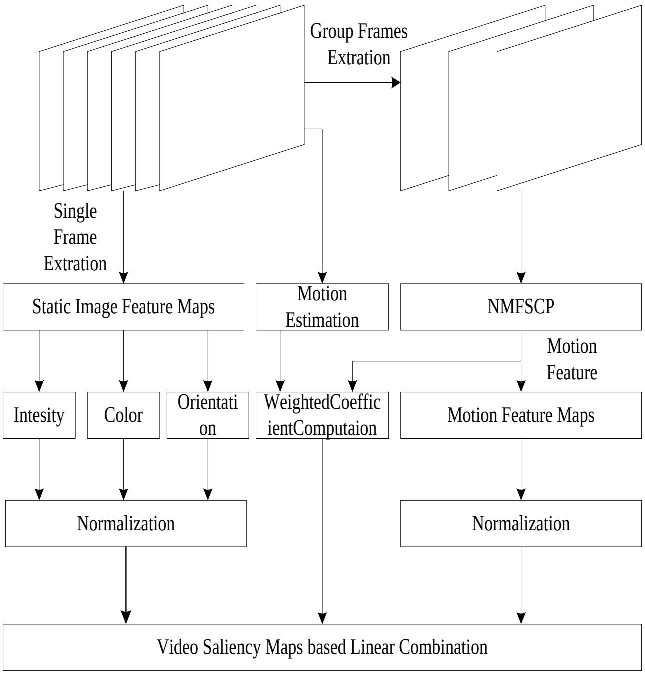

Fig.2 structure and flowchart of visual attention model proposed by this paper

There are three main parts to be processed video data in this model and there are processed parallel.

(1)Extract the static image feature of each frame with intensity, color and orientation by Itti's model of 1998 so we can obtain the saliency maps of intensity, color and orientation by $\bar{I}$,$\bar{C}$, $\bar{O}$ after normalization of Itti's model.

(2)Motion feature of video is extracted by the NMFSCP algorithm and we can obtain the motion feature image, then using Itti's normalization function we obtain the saliency map of motion feature by $\bar{M}$.

(3)Motion estimation algorithm of PMVFAST[9] is used to calculate weighted coefficient of linear combination. Block of 8*8 pixels is considered as base unit of motion estimation. Alphabet K denotes the number of block of video frame. The value of K can be obtained easily by formula (8).

$$K = \frac{M \times N}{8 \times 8} \qquad (8)$$

Next we statistic the ratio of MV of motion feature area extracted by NMFSCP in all block of frame, shown as:

$$\alpha_{MV} = \frac{\sum_{m=0}^{K-1} MV_m}{\sum_{i=0}^{M/8-1} \sum_{j=0}^{N/8-1} MV_{ij}} \qquad (9)$$

$MV = |MV_x| + |MV_y|$, MV_x and MV_y are the motion vectors of frame block in row and column direction respectively.

Last saliency map of video can be output with formula (10). It expressed that the bigger motion area and the faster motion velocity are, the larger coefficient α_M is, motion part of video will stand out, this result obey the principle of HVS that human can focus on the motion object with high speed, and it verified the efficiency of our visual attention model.

$$\begin{cases} S = \alpha_I \bar{I} + \alpha_c \bar{C} + \alpha_o \bar{O} + \alpha_M \bar{M} \\ \alpha_I + \alpha_c + \alpha_o + \alpha_M = 1 \\ \alpha_I = \alpha_c = \alpha_o \\ \alpha_M = \alpha_{MV} = \dfrac{\sum_{m=0}^{K-1} MV_m}{\sum_{i=0}^{M/8-1} \sum_{j=0}^{N/8-1} MV_{ij}} \end{cases} \qquad (10)$$

V. SIMULATION RESULT

Simulation environment is as follow: input video is hall_qcif.yuv, resolution of video is 176*144 and fps is 25fps, total is 150 frames. Experiments take one key frame every 10 frames beginning from 6th frame. We give our experiment result, there are Saliency Maps and tracing image of 16th, 26th and 36th frame in Fig.3 (a), (b), (c).Saliency map of key frame show the motion feature of whole image, with the moving of man in video saliency map change the position of conspicuity, but take effect of color, intensity and orientation into consideration, saliency map still include the static image element factor, our visual attention model did not reflect the motion feature purely. Tracing image is gotten by analyzing the saliency map to obtain the motion area. Experiment results show that our model give the good motion feature of video and tracing result.

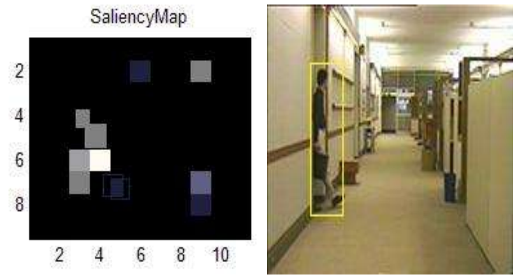

Fig.3 (a) Saliency Map and tracing result of 16thframe

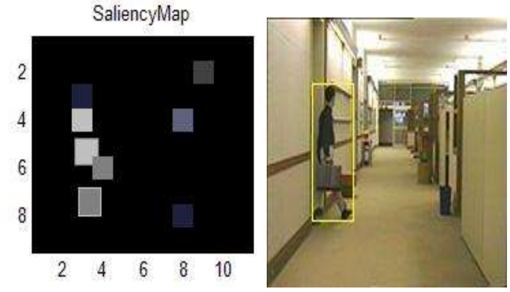

Fig.3 (b) Saliency Map and tracing result of 26thframe

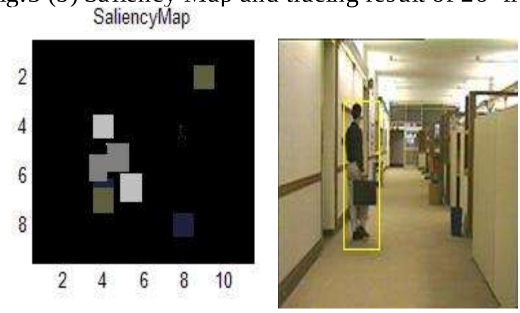

Fig.3 (c) Saliency Map and tracing result of 36thframe

VI. CONCLUSIONS

This paper presents a novel visual attention model for video based on NMFSCP. Four types of visual attention features used in the paper are intensity, color, orientation and motion. Motion feature of video key frame was extracted by algorithm of NMFCP. Intensity, color and orientation features were gotten by the Itti visual attention model. Motion vectors are calculated by PMVFAST of motion estimation algorithm and utilized to obtain the weighted coefficient of linear combination. Four features were combined with unequal linear efficient according to the ratio of motion block of motion feature in video frame. Experiments result of motion feature extracting, saliency map generation and tracing image shows the efficiency of our model. In the future we try to utilize incremental non-negative matrix factorization to speed the motion feature extracting to improve real-time capability of algorithm and we will try to apply our visual attention model into video tracing and image recognition.

ACKNOWLEDGMENT

This paper was supported by the Industrial Public Relation Project of Shaanxi Technology Committee (2010K01-077), Fundamental Research Funds for the Central Universities and 111 Project (B08038).

REFERENCES

[1] C. Koch, S. Ullman. "Shifts in Selection in Visual Attention: toward the Underlying Neural Circuitry", Human Neurobiology, 1985, 4(4), pp.219-227.

[2] Itti, L., Koch, C., and Niebur, E. 'A model of Saliency-based visual attention for rapid scene analysis ', IEEE Trans. Pattern Anal. Mach. Intell., 1998, 20, (11), pp. 1254–1259.

[3] Zhang Longfei, Cao Yuanda, Ding Gangyi, Wang Yong. A Computable Visual Attention Model for Video Skimming. In Proceedings of ISM'2008. pp.667~672.

[4] Junyong You, Jari Korhonen, Andrew Perkis. Attention modeling for video quality assessment: Balancing global quality and local quality. In Proceedings of ICME'2010. pp.914~919.

[5] Hua Zhang, Xiang Tian, Yaowu Chen: Video image assessment with a distortion-weighing spatiotemporal visual attention model. Multimedia Tools Appl. 52(1): 221-233 (2011).

[6] D. Walther, C. Koch, "Modeling Attention to Salient Proto-objects",Neural Networks, 2006, 19, pp. 1395-1407.

[7] J.Y. You, G.Z. Liu, H.L. Li, "A Novel Attention Model and Its Application in Video Analysis", Applied Mathematics and Computation,2007, 185, pp. 963–975.

[8] P. O. Hoyer. Non-negative matrix factorization with sparseness constraints. J. of Mach. Learning Res., 5:1457–1469, 2004.

[9] Alexis M.Tourapis,O.C.Au, M.L.Liou. Predictive motion vector field adaptive search technique-enhancing block based motion estimation1 Visual Communications and Image Processing 2001 (VCIP2001) , San Jose , CA , 2001.