DISTANCE MAP OF VARIOUS WEIGHTS: A NEW FEATURE FOR ADAPTIVE OBJECT TRACKING

Lin Ma[1], *Junliang Xing*[1], *Xiaoqin Zhang*[2], *Weiming Hu*[1]

1. Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China
2. College of Mathematics & Information Science, Wenzhou University, China
E-mail: lin.ma@ia.ac.cn

ABSTRACT

In this paper, we propose a new feature, Distance Map of Various Weights (DMVW) based on distances between rows' textures, to perform tracking. The proposed new feature provides an effective object appearance model which is both illumination-invariant and robust to occlusion. We also develop a 2D PCA based method to effectively evaluate the new feature. We demonstrate the validity of the rows' or column's weights in computing 2D PCA subspaces. To balance the importance of local and global information, we define a coefficient to revise the locality extent of the proposed feature. A new method based on entropy of candidate state evaluation is proposed to select the most discriminative coefficient. Experimental results on challenging video sequences demonstrated the effectiveness of our method.

Index Terms— tracking, particle filter, 2D PCA

1. INTRODUCTION

Object tracking is an important research field in computer vision and is able to provide the basis for higher level process such as motion understanding and human-machine interaction. To obtain promising tracking results, various features are adopted, such as sparse representation [1, 2], fragment [3], Harr-like feature [4], super pixel [5], etc. Histogram is a robust feature in representing the object appearance [6]. The spatial information, however, is lost in color histogram which makes the tracking easily influenced by camouflage objects nearby. To represent the relations of the points in different spatial positions, subspace is adopted [7]. Usually, object tracking is performed based on single feature [6, 7]. In many conditions, however, using multiple features makes the tracking results more robust. The VTD [8] algorithm decomposes object appearances to various components to tackle different tracking challenges. Generally traditional methods do not consider the components' distances as feature and fail to tackle some challenges, e.g. drastic light variation. Distance based features, however, is able to tackle the light variation problem etc. effectively. Some researchers perform tracking with heat kernel and the distances between object sub image pixels are utilized to represent the relations between the pixels [10]. Distances between samples are also important in representing the relations of samples. When obtaining the distances between samples, MDS (Multidimensional Scaling) is able to reconstruct the samples (with translation invariance, etc.) [9]. Distance based discriminative models, i.e. distances between foreground samples and background samples, are also proposed to determine object states [14, 15, 18].

The work presented here focuses on distance based features. Distances between pixels are also considered in [10]. However, different from [10], our feature need not perform spectral decomposition on high dimension matrix and thus is able to be obtained efficiently. The work in [14, 15, 18] also utilize distance information. But these work represent the relations between samples, while our work considers the relations between sample' components. Thus, our method is able to tackle the light variations and occlusions more effectively.

To represent the local appearance effectively, we divide the object appearance into patches. For each patch, for efficiency we only utilize the distances between row textures to construct the DMVW feature. A coefficient is defined to represent the locality extent of the DMVW feature. The entropies of a set of candidate states' likelihoods are able to represent the discriminative ability of the feature effectively. Thus, we utilize the entropies to select the optimal locality extent coefficient. 2D PCA is adopted for efficient evaluation of the object state. The key contributions of this paper are as follows.

(1) We propose a new feature, DMVW, which utilizes the benefit of distance information and represents effectively the local and global information of the object appearance.
(2) We use the entropy of candidate states' evaluations to select the optimal locality extent coefficient. The entropy represents the discriminative ability of the feature effectively.
(3) We demonstrate the validity of the rows' or column's weights in computing 2D PCA subspaces.

The rest of our paper is organized as follows: Section 2 gives a short discuss of the particle filter we use in this paper. In Section 3, we construct and evaluate the proposed DMVW feature. Experimental results are shown in Section 4 and conclusion and future work are made in Section 5.

2. PARTICLE FILTER BASED FRAMEWORK

In this paper, we adopt particle filter as the framework for tracking. Particle filter is formed based on Bayesian

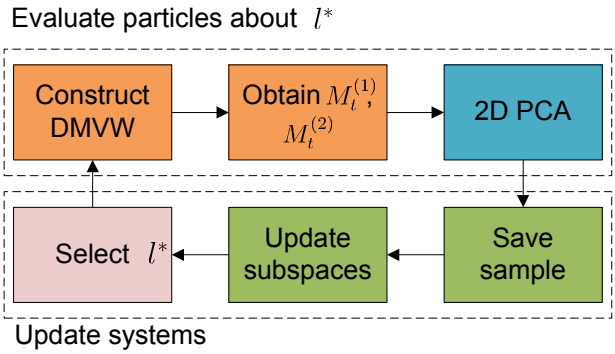


Fig. 1: The flowchart of our system. Our system is under the particle filter framework and contains two main parts. First, evaluate the particles with the most discriminative feature (about l^*). Second, update the subspaces about each l and select l^*.

formula [13]. Let X_t^x, X_t^y, X_t^{sx}, X_t^{sy} be the x, y position and width and height scales of the object respectively, then we define the object state in frame t with a vector $X_t = \{X_t^x, X_t^y, X_t^{sx}, X_t^{sy}\}$. Given the observation sequence $O_{1:t+1}$, the object state's posteriori probability density is defined as

$$p(X_{t+1}|O_{1:t+1}) \propto p(O_{t+1}|X_{t+1}) \int p(X_{t+1}|X_t) p(X_t|O_{1:t}) dX_t. \quad (1)$$

We warp linearly the sub image specified by X_t to a normalized 32×32 sub image.

In each frame, we first sample some particles according to (1), and then select the particle with the largest likelihood $\pi_t = p(O_t|X_t)$ as the optimal state. The tracking is performed on gray scale images. Detailed presentation of particle filter can be found in [13].

3. DISTANCE MAP OF VARIOUS WEIGHTS

To represent the local information of the target more effectively, we divide the normalized sub image into 6×6 patches. The neighbor patches are overlapped by each other. We obtain DMVW feature for each patch independently. The features are evaluated with 2D PCA method.

3.1. Constructing DMVW

We represent each patch's appearance with distance information which represents the relations between object components effectively and is robust to light variation etc. To reduce the dimensionality of the feature, we only consider the distances between row textures. For one patch, let N_R be the number of rows and r_i be row i's texture (Fig. 2). Then the relation of r_i and r_j is defined as

$$q^l(i,j) = exp(-\frac{||r_i - r_j||_2}{N_R}) \frac{|i-j|^l}{\bar{i}}, \quad (2)$$

where $\bar{i} = \sum_j |i-j|^l$, and $l \in R$ is a coefficient representing the feature's locality extent. Enlarging l increases the weights of $q^l(i,j)$ with large $|i-j|$ and then makes the feature represent more global information, and vice verse. Then for the patch we form the DMVW feature $\tilde{q}^l = [q^l(0,0), q^l(0,1), ..., q^l(5,5)]^T$.

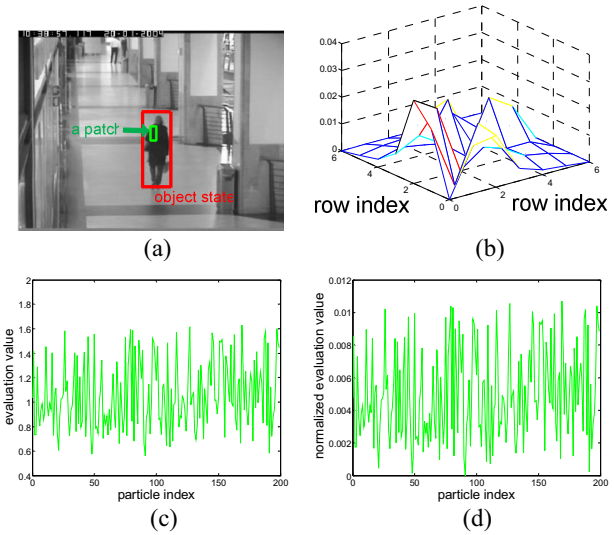


Fig. 2: Object appearance and evaluations about l=1. (b) The DMVW (before being unfolded to vector) of the specific patch in (a). (c) Particle evaluations. (d) The normalized evaluations of (c).

We denote patch (i,j)'s feature as $\tilde{q}_{i,j}^l$, $i,j = 0,\ldots,5$. Then we obtain two matrices $M^{l,(2)} = [\tilde{q}_{0,0}^l, \tilde{q}_{0,1}^l, ..., \tilde{q}_{5,5}^l]$ and $M^{l,(1)} = {M^{l,(2)}}^T$ to represent the object appearance. For conciseness, we omit the script of l in symbols normally. The appearance is evaluated with 2D PCA.

3.2. Evaluating object appearance with 2D PCA

2D PCA is an efficient model in representing the relations between matrix columns. Thus, we adopt 2D PCA [11, 12] to compute the subspaces U_1 and U_2 of $M_t^{(1)}$ and $M_t^{(2)}$ respectively. For different l, 2D PCA is performed separately. We define the mean of $M_t^{(i)}$, $t = 0, \cdots, n+m$ as

$$\overline{M}_{m+n}^{(i)} = s\overline{M}_n^{(i)} + (1-s)\widetilde{M}^{(i)} \quad , i = 1, 2, \quad (3)$$

where $\widetilde{M}^{(i)}$ is the mean of $M_t^{(i)}$, $t = n+1, \cdots, n+m$, s is a constant which defines the weight of historic information. The covariance matrix $D_n^{(i)}$ of $M_t^{(i)}$, $t = 0, \cdots, n$ is computed in the similar way to the mean matrix. Then U_i is spanned by $D_n^{(i)}$'s eigenvectors.

Different patches have different confidences. Based on the confidences, we define patch (i,j)'s weights $w_0^{i,j}$ and $w_1^{i,j}$ in updating subspaces and evaluating particles respectively (Section 3.3). Let $W_t = \text{Diag}\left(w_1^{0,0}, w_1^{1,0}, ..., w_1^{5,5}\right)$ be the weight matrix in time t for evaluating the particles. The DMVW of different l have different abilities to discriminate the object from the background. During tracking, we select the most discriminative coefficient l^* to evaluate the particles. The distances between the object appearance specified by X_t and U_i, $i = 1, 2$ are defined as

$$E_{t,1} = \left\| W_t((M_t^{(1)} - \overline{M}_t^{(1)}) - U_1 U_1^T (M_t^{(1)} - \overline{M}_t^{(1)})) \right\|_2, \quad (4)$$

$$E_{t,2} = \left\| ((M_t^{(2)} - \overline{M}_t^{(2)}) - U_2 U_2^T (M_t^{(2)} - \overline{M}_t^{(2)})) W_t \right\|_2 . \quad (5)$$

Then the likelihood of X_t is

$$p(O_t|X_t) \propto \sum\nolimits_{i=1}^{2} \exp(-E_{t,i}). \quad (6)$$

The process of our method is shown in Algorithm 1. Next we show the validity of the weights' definition. That is, the patches with larger weights play more important roles in computing U_1 and U_2.

Proof. In this proof, we only consider offline conditions and assume that there are 5 frames. Let the 6×6 patches' appearance vectors and weights in frame *t* be $v'_{t,i}$ and w'_i (assume invariant), $i = 0, 1, \cdots, 35$ respectively. We define

$$M_t^{(2)} - \overline{M}_t^{(2)} = [v'_{t,0}, v'_{t,1}, \cdots, v'_{t,35}]. \quad (7)$$

Then U_2 is spanned by the eigenvectors of

$$D_4^{(2)} = C_1 \sum\nolimits_{t=0}^{4} \sum\nolimits_{i=0}^{35} w'^2_i v'_{t,i} v'^T_{t,i}, \quad (8)$$

where C_1 is a constant. From (8), we see that patches with larger weights can be considered to have more samples, and then are more important in computing U_2. So we have demonstrated the validity of patches' weights in computing U_2.

Then we show the validity of weights $w'_i, i = 0, 1, \cdots, 35$ in computing U_1. According to $M_t^{(1)}$'s definition, we obtain

$$M_t^{(1)} - \overline{M}_t^{(1)} = [v'_{t,0}, v'_{t,1}, \cdots, v'_{t,35}]^T. \quad (9)$$

U_1 factually is the subspace of the column vectors of $\mathrm{Diag}\,(w'_0, ..., w'_{35})\,(M_t^{(1)} - \overline{M}_t^{(1)})$, $t = 0, \cdots, 4$. The data of the *i*-th patch vector corresponds to the *i*-th entries of the column vectors. Larger w'_i tends to enlarge the data variance and then the subspace is more likely to be influenced by the *i*-th entries. Thus, we have shown the validity of w'_i, $i = 0, 1, \cdots, 35$ in computing U_1. So we have demonstrated the validity of w'_i, $i = 0, 1, \cdots, 35$ in computing U_1 and U_2 (in offline conditions). □

3.3. Weights of patch

Different patches generally vary to different extents, and the variation condition is able to represent the confidence of the patch. Thus, we define the patches' weights according to the variation conditions (probability density). For different *l*, the weights are defined separately. The patch variation in the previous frame can represent the occlusion condition effectively. For patch (*i*, *j*), let $\bar{q}_{i,j}$ and $v_{i,j}$ be the mean and variance of $\widetilde{q}_{i,j}$ of all frames. Then given the patch vector $q_{i,j}$ in frame *t*, the patch's weight in frame *t*+1 is defined as

$$w_1^{i,j} \propto 1/(\sqrt{2\pi v_{i,j}}) \exp\{-||q_{i,j} - \bar{q}_{i,j}||_2^2/(2v_{i,j})\}. \quad (10)$$

When updating the system, we need to store new samples. Same with the evaluation of particles, we check the current frame's sample only based on DMVW of l^*. For patch (*i*, *j*), if $||\widetilde{q}_{i,j} - \bar{q}_{i,j}||_2^2 < \alpha v_{i,j}$, we consider patch (*i*, *j*) is valid, where α is a constant. If the number of valid patches (0∼36) is larger than a threshold, we consider the current frame is valid and store the current frame's DMVW of all *l* for system updating, otherwise the current sample is dropped. The system is updated every 5 valid frames.

We define the patches' weights in computing U_1 and U_2 according to the stored samples. For different *l*, the patch's weights are defined separately. Assume patch (*i*,*j*)'s weights of stored sample *k* is $w'^{i,j}_k$, then we define $w_0^{i,j}$ as the mean of $w'^{i,j}_k$, $k = 0, \cdots, 4$.

3.4. Selecting l^*

We utilize particles' evaluations to select l^*. Generally, better *l* makes the evaluations of particles more different from each other. Entropy is able to represent the difference between particles' evaluations effectively. Thus, we use entropy to select l^*. We define $g^{l,i}$ as the evaluation of particle *i* about *l*, N_p as the particle number and g^l_{max} and g^l_{min} as the largest and the smallest values of $g^{l,i}$, $i = 0, \ldots, N_p - 1$ respectively. Let $\widetilde{g}^{l,i} = g^{l,i} - g^l_{min}$, and $\widetilde{g}^{l,i}$, $i = 0, \ldots, N_p - 1$ is normalized with the sum to 1 constraint. Then $\widetilde{g}^{l,i}$ is able to be considered as probability and fulfills the condition of entropy. Sometimes g^l_{min} is large compared with $g^l_{max} - g^l_{min}$, which makes the entropy based evaluation of *l* not effective. Thus, we use $\widetilde{g}^{l,i}$ not $g^{l,i}$ to compute the entropy about *l* (Fig. 2(c)(d))

$$H^l = -\sum\nolimits_i \widetilde{g}^{l,i} \log(\widetilde{g}^{l,i}). \quad (11)$$

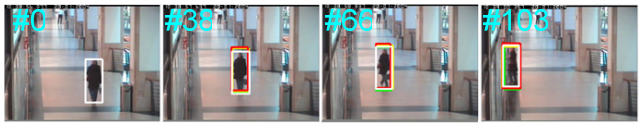
For each valid frame, we save the entropy about each *l*. Every 5 valid frames, we compute the mean $\widetilde{H}^l$ of the 5 entropies about *l*. Normally the smaller $\widetilde{H}^l$ is, $\widetilde{g}^{l,i}$, $i = 0, \ldots, N_p - 1$ are more different from each other, and the particles are more easily discriminated from each other, and vice verse. Thus we select the *l* with smallest $\widetilde{H}^l$ as l^*.

Algorithm 1 The process of our method.

Tracking: For each frame *t*, do:
1) Sample particles with particle filter.
2) Find the optimal particle with (6).
Updating:
1) Check if the current frame is valid with DMVW of l^*.
2) If current frame is valid, save the frame's features of all l. Update W_t.
3) Select new l^* with (11) every 5 valid frames.
4) Update the 2D PCA subspaces corresponding to DMVW of each l every 5 valid frames.

4. EXPERIMENTS

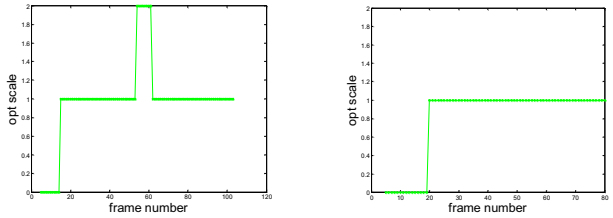
We implemented our method in C++ and evaluated it on 5 video sequences involving multiple challenges, such as light variation, occlusion, etc. The experiments were conducted on a PC with a 2.53 GHz Intel CPU with 2GB RAM. For each experiment of our method, the object state in the first frame was manually set, and the system was updated every 5 saved samples. We created 200 particles per frame during tracking and defined *s* as 0.95-0.98. We defined three kinds of locality extent coefficients, i.e. *l*=0,1,2. The running time of our method



(a) Walking woman1. Pose variation.

(b) Walking woman2. Camouflage.

Fig. 3: Comparison between l=0, 1, 2 and our method.



(a) walking woman1 (b) walking woman2

Fig. 4: Selection of optimal l, i.e. l^* .

Table 1: Average errors about the 4 methods in Fig. 3

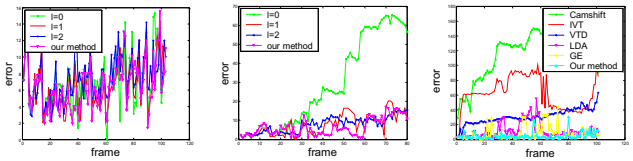
	l=0	l=1	l=2	Our method
Fig. 3(a)	7.0	6.7	7.5	**6.5**
Fig. 3(b)	28.8	7.6	7.8	**6.3**

was around 1.4 sec/frame. The test videos we used were from [16] and [17]. We adopted the Euclidean distance between the object bounding box's center and the ground truth to represent the tracking error.

In Fig. 3, we tested our method in using different locality extent coefficients, i.e. l=0,1,2. Smaller l represented the local information effectively, while larger l represented the global information effectively. By selecting the most discriminative l, our method was able to discriminate the foreground from the background more effectively, and thus obtained more accurate object states. Fig. 4 showed the selections of the optimal l in the two videos of Fig. 3. The average errors and the error maps of the 4 methods were shown in Tab. 1 and Fig. 6 respectively.

In Fig. 5, we compared our method with other 5 methods which adopted different appearance models: Camshift [6], IVT [7], VTD [8], LDA and GE [15]. In Fig. 5, the object experienced drastic light variation. The color histogram of Camshift changed largely, which made Camshift influenced by nearby similar colors and lose tracking. IVT represented object appearances with linear subspaces. Due to the large light variation, the subspaces failed to represent the distributions of the samples accurately and made the method gradually drift away. VTD utilized various features to represent the object. When light changed drastically and also similar color area existed nearby, the combined features were also influenced largely which made VTD not robust any more. LDA

Fig. 5: Head. Comparison with 5 methods. Light variation.



(a) walking woman1 (b) walking woman2 (c) head

Fig. 6: Error maps.

Fig. 7: Occlusion conditions in two videos about 2 walkers.

and GE were able to discriminate object from background effectively. But they failed to represent the distances between object components which were robust to light variation. Thus, the two methods were not able to track the object accurately. In contrast, our proposed DMVW feature took advantage of the robust distance information between row textures in light variation condition. Thus, our method was able to track the object robustly. The error maps of the 6 methods were shown in Fig. 6, and the average errors of Camshift, IVT, VTD, LDA, GE and our method were 111.8, 62.3, 30.9, 11.4, 8.4 and 4.6 respectively. Moreover, as Fig. 7 showed, by setting the patches' weights according to occlusion conditions, our method also tackled the occlusions effectively.

5. CONCLUSION AND FUTURE WORK

This paper has proposed a new feature DMVW to perform object tracking. The DMVW feature took advantage of the distance information between different object components, and was robust to light variation, etc. 2D PCA was adopted to obtain the optimal state and we demonstrated the validity of patches' weights in updating 2D PCA subspaces. We also proposed a novel method to select the most discriminative locality extent coefficient. Experiments showed our method's effectiveness. In the future, we will continue to research more effective features based on distance information. (This work is partly supported by NSFC (Grant No. 60935002, 61100099, 61100147), the National 863 High-Tech R&D Program of China (Grant No. 2012AA012504), the Natural Science Foundation of Beijing (Grant No. 4121003), and The Project Supported by Guangdong Natural Science Foundation (Grant No. S2012020011081), NSF of Zhejiang Province (Grant No. LY12F03016).)

6. REFERENCES

[1] B. Liu, J. Huang, L. Yang, and C. Kulikowsk, "Robust tracking using local sparse appearance model and k-selection", *CVPR*, 2011.

[2] X. Mei, and H. Ling, "Robust visual tracking and vehicle classification via sparse representation", *TPAMI*, 33(11), pp. 2259-2272, 2011.

[3] A. Adam, E. Rivlin, and I. Shimshoni, "Robust fragments-based tracking using the integral histogram", *CVPR*, 2006.

[4] B. Babenko, M. Yang, and S. Belongie, "Visual tracking with online multiple instance learning", *CVPR*, 2009.

[5] S. Oron, A. Bar-Hillel, D. Levi, and S. Avidan, "Locally orderless tracking", *CVPR*, 2012.

[6] G. R. Bradski, "Computer vision face tracking for use in a perceptual user interface", *Intel Tech. J.*, 2(Q2), 1998.

[7] D.A. Ross, J. Lim, R.S. Lin, and M.H. Yang, "Incremental learning for robust visual tracking", *IJCV*, 77(1), pp. 125-141, 2008.

[8] J. Kwon and K. M. Lee. "Visual tracking decomposition", *CVPR*, 2010.

[9] T.F. Cox and M.A.A. Cox, "Multidimensional scaling", *Chapman & Hall*, London, 1994.

[10] X. Li, W. Hu, H. Wang, and Z. Zhang, "Robust object tracking using a spatial pyramid heat kernel structural information representation", *Neurocomputing*, 73(16-18), pp. 3179-3190, 2010.

[11] J. Yang, D. Zhang, A. F. Frangi, and J. Y. Yang, "Two-dimensional PCA: a new approach to appearance-based face representation and recognition", *TPAMI*, 26(1), pp. 131-137, 2004.

[12] T. Wang, I. Y. H. Gu, and P. F. Shi, "Object tracking using incremental 2D-PCA learning and ML estimation", *ICASSP*, 2007.

[13] M. Isard and A. Blake, "CONDENSATION - Conditional density propagation for visual tracking", *IJCV*, 1(29), pp. 5-28, 1998.

[14] X.Q. Zhang, W.M. Hu, S. Maybank, and X. Li, "Graph based discriminative learning for robust and efficient object tracking", *ICCV*, 2007.

[15] L. Ma, W.M. Hu, and X.Q. Zhang, "Multiple sample group pairs' graph embedding for tracking", *ICIP*, 2012

[16] http://homepages.inf.ed.ac.uk/rbf/CAVIAR/

[17] L. Ivan, M. Marcin, S. Cordelia, and R. Benjamin, "Learning realistic human actions from movies", *CVPR*, 2008.

[18] G. Li, D. Liang, Q. Huang, S. Jiang, and W. Gao, "Object tracking using incremental 2D-LDA learning and Bayes inference", *ICIP*, 2008.