GRAYSCALE-THERMAL TRACKING VIA CANONICAL CORRELATION ANALYSIS BASED INVERSE SPARSE REPRESENTATION

Wan Ding[1], *Bin Kang*[2], *Quan Zhou*[1], *Min Lin*[1], *Suofei Zhang*[2]

1. College of Communication and Information Engineering,
Nanjing University of Posts and Telecommunications, Nanjing, China
2. College of Internet of Things,
Nanjing University of Posts and Telecommunications, Nanjing, China

ABSTRACT

The grayscale-thermal tracking has attracted increasing attention due to the fact that it can make thermal information complement with grayscale information. Since there exists a large gap between the grayscale and the thermal video sequences, how to exploit the intrinsic relation between the grayscale and the thermal targets has become the key point. To address this issue, in this paper, we propose an inverse sparse representation based framework for the grayscale-thermal tracking, in which a canonical correlation analysis based inverse sparse representation model is adopted to jointly encode the target candidates in the grayscale and the thermal video sequences. The target coding process can explore the similarity between the grayscale and the thermal appearance in a common subspace, which can highlight the useful and discriminative information in both grayscale and thermal targets. The experiments on OSU-CT dataset can illustrate the promising performance of our tracking framework.

Index Terms— Grayscale-thermal tracking, Inverse sparse representation, Canonical correlation analysis

1. INTRODUCTION

Visual tracking plays a very important role in computer vision with many applications, such as video analysis, vehicle navigation and human-computer interaction. Although significant progress has been made recently, it remains very challenging for visual tracking under bad weather, such as smog and raining, because the visible spectrum camera can only collect limited light, causing foreground target and the background difficult to be discriminated.

With the rapid development of multimedia and internet of things, thermal infrared camera has become economically affordable. This kind of camera can capture the thermal infrared radiation emitted by subjects with a temperature above absolute zero, which is good for night surveillance. Effectively combining visible spectrum camera with thermal infrared camera has two advantages: 1) Thermal infrared camera is robust to the illumination changes, which can provide complement to visible spectrum data obtained from poor light condition. 2) The gray feature in visible spectrum camera would contribute to solving the crossover problem in thermal infrared camera based object detection. In this context, the grayscale-thermal tracking is considered as an effective method to overcome the bad weather challenging in visual tracking [1, 2].

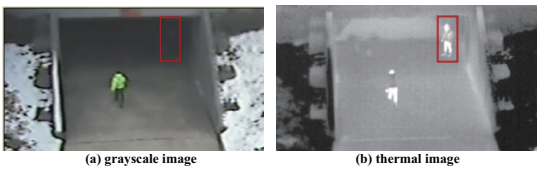


Fig. 1. An example of grayscale and thermal video pairs. In (a), it is difficult to discriminate the target from background. The grayscale-thermal tracking aims to design the appearance model to make thermal and grayscale information complement with each other.

The grayscale and thermal video sequences are obtained in pairs in grayscale-thermal tracking (see Fig. 1 as example). Based on the video pairs, designing the appearance model is a tough work because it not only requires to bridge the image gap between the grayscale and the thermal video sequences, but also asks to resist the data bias in the grayscale or thermal video sequences. Due to the successful application of sparse representation in multi-view face recognition [3], sparse representation has become a useful tool to overcome the limitation in grayscale-thermal tracking. Different from traditional sparse representation based visual tracking [4], the grayscale-thermal sparse representation based tracking has to adaptively fusing the sparse representations of grayscale and thermal appearance to guarantee that both the grayscale and the thermal video sequences can give good tracking performance. For

Thanks to NSFC (Nos. 61801242,61876093,61771258 and 61707252), NSFJS (Nos. BK20170915, BK20181393) and NUPT program (NY218075).

practical application, Li [2] proposed a collaborative sparse representation based appearance model for grayscale-thermal tracking, in which the multi-model fusion and the model reliability estimating are integrated into a unified optimization problem. Since the collaborative sparse representation model could not explore the similarity between the grayscale and the thermal video sequences, it may not make sure that the same target in the grayscale and the thermal image pairs can obtain similar sparse representation results when facing data bias. To enhance the tracking performance of [2], Li [5] introduced target local information in the collaborative sparse representation model. Since this method could not use background context to enhance the difference between similar targets, it may cause tracking drift in background clutter.

Inverse sparse representation is an extension of sparse representation, which can be used as a feature coding method to separate the target from background [6, 7]. Inspired by this method, we propose an inverse sparse representation based tracking framework for robust grayscale-thermal tracking, in which the canonical correlation analysis (CCA) and inverse sparse representation are firstly integrated into a joint optimization problem to encode the target candidates. Then, the target candidate codes are put in SVM method for grayscale-thermal tracking. The advantages of joint optimization based observation coding are two-folds: 1) The discriminative information of targets in both grayscale and thermal images can be highlighted through using CCA to explore the target similarity between grayscale and the thermal video sequences. 2) Due to the robustness of inverse sparse representation, the observation codes are robust to the data bias. To enhance the tracking speed, we also propose an alternating reconstruction method to solve the joint optimization problem.

2. TRADITIONAL INVERSE SPARSE REPRESENTATION

In traditional sparse representation based visual tracking, sparse representation aims to use an dictionary to sparsely represent the target candidates. Different from sparse representation method, inverse sparse representation is aimed to use target candidates to sparsely represent the target dictionary. Since the target dictionary is composed of the target and the background templates, using target candidates to inversely represent the target dictionary can indicate the similarity of target candidates to target templates and background templates. In this way, the inverse sparse representation can use both the target and the background information to yield the discriminative codes to represent the target candidates. The inverse sparse representation model is described as

$$\min_{\mathbf{U}} \|\mathbf{U}\|_1 + \|\mathbf{D} - \mathbf{Y}\mathbf{U}\|_F^2, \tag{1}$$

where $\mathbf{D} = [\mathbf{D}_P, \mathbf{D}_N]$ is the target dictionary with $\mathbf{D}_P$ and $\mathbf{D}_N$ are the positive and negative temple sub-matrix (the target and the background templates). $\mathbf{Y}$ denotes the target candidate matrix and $\mathbf{U}$ is the corresponding inverse sparse representation result. If the target is not occluded severely, equation (1) can use $\mathbf{U}$ to accurately encode target candidates in $\mathbf{Y}$. However, if the target is completely occluded or in the bad illumination scenario, equation (1) could not extract the useful information in target candidates to guarantee the sparsity in $\mathbf{U}$.

The grayscale-thermal tracking is aimed to use grayscale and thermal information to complement with each other. Since equation (1) only achieves target feature coding for a single kind of video sequence, it could not be directly used in grayscale-thermal tracking.

3. CANONICAL CORRELATION ANALYSIS BASED INVERSE SPARSE REPRESENTATION

3.1. Proposed inverse sparse coding model

In this paper, we propose a canonical correlation analysis based inverse sparse representation model to encode the target candidates. The proposed model is shown as that

$$\min_{\mathbf{U},\mathbf{P}} \sum_{i=1}^{2} \|\mathbf{P}_i^T\mathbf{D}_i - \mathbf{P}_i^T\mathbf{Y}_i\mathbf{U}_i\|_F^2 + \lambda_1\|\mathbf{U}\|_{2,1} - \lambda_2 Tr(\mathbf{P}_1^T\mathbf{Y}_1\mathbf{Y}_2^T\mathbf{P}_2), \tag{2}$$

where $\mathbf{D}_i$ $(i = 1, 2)$ means the target dictionaries obtained from the thermal and thegrayscale video sequences, respectively. $\mathbf{Y}_i$ $(i = 1, 2)$ denotes the target candidate matrices for the thermal and the grayscale video sequences, respectively. $\mathbf{U} = [\mathbf{U}_1, \mathbf{U}_2]$ is the inverse sparse representation matric for the target candidate matrices $\mathbf{Y}_1$ and $\mathbf{Y}_2$. $\mathbf{P}_i (i = 1, 2)$ denotes the projection matrices, which is updated through minimizing $-Tr(\mathbf{P}_1^T\mathbf{Y}_1\mathbf{Y}_2^T\mathbf{P}_2)$.

In computer vision, the same target can be represented by different features, such as texture, edges and so on. There exist not only the potential similarity but also a large gap between different kinds of target features. Canonical Correlation Analysis (CCA) [8] aims to exploit the intrinsic feature similarity to train the projection matrices for projecting different target features into a common subspace, in which the common and useful information of different target features can be maximized. Inspired by CCA, we integrate projection matrices updating and the inverse sparse representation into a unified optimization model, in this way, we can use $\mathbf{P}_i^T\mathbf{Y}_i$ to enforce the target similarity between the grayscale and the thermal video sequences, which can highlight the useful information and minimize the data bias in both $\mathbf{Y}_1$ and $\mathbf{Y}_2$. Based on the advantage of $\mathbf{P}_i^T\mathbf{Y}_i$, equation (2) can make sure to yield robust codes for the target candidates in both grayscale and the thermal video sequences.

3.2. Reconstruction method

Equation (2) is a non-smooth optimization problem. To solve this problem, we propose an alternating reconstruction

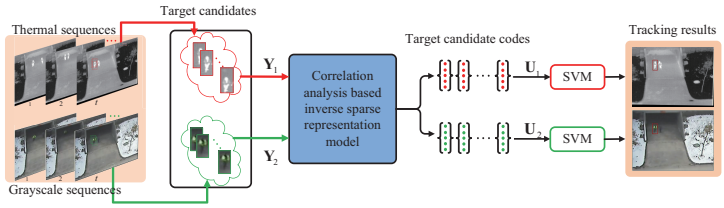


Fig. 2. The proposed tracking framework for grayscale and thermal video pairs

method, in which each iteration is divided into two steps, termed as the **P** and the **U** steps.

In **P** step, **U** is fixed, the projection matrices $\mathbf{P}_1$ and $\mathbf{P}_2$ are updated by solving the following sub-optimization problem

$$\min_{\mathbf{P}_1,\mathbf{P}_2} \sum_{i=1}^{2} \|\mathbf{P}_i^T\mathbf{D}_i - \mathbf{P}_i^T\mathbf{Y}_i\mathbf{U}_i\|_F^2 - \lambda_2 Tr(\mathbf{P}_1^T\mathbf{Y}_1\mathbf{Y}_2^T\mathbf{P}_2). \quad (3)$$

Based on the property of Frobenius norm, problem (3) can be rewritten as

$$\min_{\mathbf{P}_1,\mathbf{P}_2} \sum_{i=1}^{2} Tr(\mathbf{P}_i^T\mathbf{Q}_i\mathbf{Q}_i^T\mathbf{P}_i) - Tr(\lambda_2\mathbf{P}_1^T\mathbf{Y}_1\mathbf{Y}_2^T\mathbf{P}_2) \quad (4)$$

where $\mathbf{Q}_i = \mathbf{D}_i - \mathbf{Y}_i\mathbf{U}_i$. Using $\mathbf{P} = [\mathbf{P}_1^T, \mathbf{P}_2^T]^T$ to reformulate problem (4), it can be simplified as

$$\min_{\mathbf{P}} Tr(\mathbf{P}^T\mathbf{A}\mathbf{P}) \quad (5)$$

where $\mathbf{A} = \begin{bmatrix} \mathbf{Q}_1\mathbf{Q}_1^T & -\mathbf{Y}_1\mathbf{Y}_2^T \\ \mathbf{0} & \mathbf{Q}_2\mathbf{Q}_2^T \end{bmatrix}$, with $\mathbf{0}$ being the zero matrix. Setting the first order derivative of problem (5) to zero, we can update the projection matrix **P** through singular value decomposition.

In **U** step, **P** is fixed, and $\mathbf{U} = [\mathbf{U}_1, \mathbf{U}_2]$ is updated through solving the following optimization problem

$$\min_{\mathbf{U}} \sum_{i=1}^{2} \|\mathbf{P}_i^T\mathbf{D}_i - \mathbf{P}_i^T\mathbf{Y}_i\mathbf{U}_i\|_F^2 + \lambda_1\|\mathbf{U}\|_{2,1} \quad (6)$$

In (6), we set $\Phi(\mathbf{U}) = \sum_{i=1}^{2} \|\mathbf{P}_i^T\mathbf{D}_i - \mathbf{P}_i^T\mathbf{Y}_i\mathbf{U}_i\|_F^2$, $\Psi(\mathbf{U}) = \|\mathbf{U}\|_{2,1}$. Applying composite gradient mapping [9] to $\Phi(\mathbf{U})$ and $\Psi(\mathbf{U})$, we can obtain

$$\begin{aligned} \mathbf{U}^{k+1} = &\min_{\mathbf{U}} \Phi(\mathbf{V}^k) + < \nabla\Phi(\mathbf{V}^k), \mathbf{U} > \\ &+ \frac{1}{2\eta}\|\mathbf{U} - \mathbf{V}^k\|_F^2 + \lambda_1\Psi(\mathbf{U}) \end{aligned} \quad (7)$$

where η is the step size parameter. Inspired by [10], the solution of (7) is given by

$$\mathbf{U}^{k+1/2} = \mathbf{V}^k - \eta\nabla\Phi(\mathbf{V}^k) \quad (8)$$

where $[\nabla\Phi(\mathbf{V}^k)]_i = -(\mathbf{P}_i^T\mathbf{Y}_i)^T(\mathbf{P}_i^T\mathbf{D}_i - \mathbf{P}_i^T\mathbf{Y}_i\mathbf{U}_i)$ $(i = 1, 2)$

Based on (8), **U** is finally updated by

$$[\mathbf{U}^{k+1}](j,:) = \left[1 - \frac{\lambda_1\eta}{\|[\mathbf{U}^{k+1/2}](j,:)\|_2}\right]_+ [\mathbf{U}^{k+1/2}](j,:) \quad (9)$$

where $[\mathbf{U}^{k+1}](j,:)$ represents the j-th row in matrix $\mathbf{U}^{k+1}$ and $[\cdot]_+$ is the scalar operator. Assuming there exist a scalar a, it is defined that $[a]_+ = \max\{0, a\}$.

4. INVERSE SPARSE REPRESENTATION BASED TRACKING FRAMEWORK

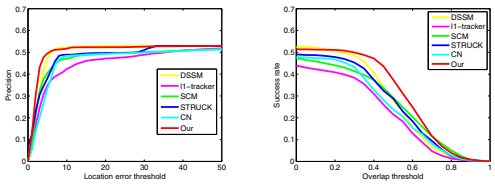
In this section, we will illustrate how to use the correlation analysis based inverse sparse representation model to achieve visual tracking. The online tracking framework is shown in Fig.2.

At time t, we firstly adopt particle filter [11] to yield the target candidate matrices $\mathbf{Y}_1$ and $\mathbf{Y}_2$. Then we use the proposed inverse sparse representation model (equation (2)) to jointly estimate the target candidate codes $\mathbf{U}_1$ and $\mathbf{U}_2$ for $\mathbf{Y}_1$ and $\mathbf{Y}_2$. Since $\mathbf{U}_1$ and $\mathbf{U}_2$ are estimated through exploring the similarity between the grayscale and the thermal video sequences, putting them into SVM can make it easy to discriminate the best target from the target candidates in both the grayscale and the thermal video sequences. To avoid miscoding, the target dictionaries $\mathbf{D}_i$ $(i = 1, 2)$ are online updated in a manner similar to [6]. Inspired by [12], the SVM is pre-trained by using the positive and negative sample codes (the positive and negative training samples are encoded by the proposed inverse sparse representation model) in the first 10 frames, and SVM is online updated at every 50 frames.

5. EXPERIMENTS

OSU-CT is a public dataset [13] for testing the grayscale-thermal tracking performance, in which it contains 9 video pairs with challenging factors such as: bad illumination, occlusion etc. Here we use this dataset to carry out the experiments. We compare the proposed method with 7 state-of-the-art methods, namely L1-tracker [4], SCM [14], Struck [15], CN [16], DSSM [7], L1-PF [17] and GTOT [5]. In those methods, L1-tracker, SCM, Struck, CN and DSSM

Fig. 4. Qualitative tracking performance with different tracking methods. The first row is the grayscale video sequences, and the second row is the thermal video sequences



(a) (b)

Fig. 3. Precision and success rate performance over different tracking methods on grayscale video sequences. (a) precision plot, (b) success plot.

are grayscale tracking methods, and L1-PF and GTOT are grayscale-thermal tracking methods.

Experiment setting: In our experiments, the target dictionaries are set as $\mathbf{D}_i \in R^{256\times 300}(i = 1, 2)$, in which the number of target templates is 300 (200 for foreground templates and 100 for background templates). The number of target candidates in each frame is 600. The parameters in the reconstruction method are empirically set as $\eta = 0.01$, $\lambda_1 = 0.1$ and $\lambda_2 = 1$.

In the experiments, we firstly give the quantitative evaluation of 6 tracking methods (see Fig. 3). SCM, Struck, CN, l1-tracker and DSSM are well-known grayscale based tracking methods in which they only use grayscale information to carry out visual tracking. By comparison, our method uses grayscale-thermal video pairs to carry out visual tracking. This experiment aims to test whether the grayscale-thermal method can enhance the grayscale tracking performance with the help of thermal information in the challenging video sequences. The quantitative measurements used for this test include the precision plot and the success plot [2]. These two measurements are often used to evaluate the overall tracking performance. From Fig. 3 we can clearly see that our method can give the best tracking performance in the precision and success plots.

Next, we quantitatively compare our method with the well-known grayscale-thermal tracking methods (see Table 1). In this test we use overlap rate as the objective measurement, which is defined as $\frac{area(B_T \cap B_G)}{area(B_T \cup B_G)}$, where B_T and B_G are the tracked bounding box of each frame and the corresponding ground truth, respectively. From Table 1 we can clearly see that the average overlap rate of out method in grayscale and thermal video sequences are 0.60 and 0.61, respectively, which is almost 8% higher than GTOT method.

Table 1. Average overlapping rate performance over different grayscale-thermal trackers. The best results are denoted as red

	grayscale performance			thermal performance		
video pairs	L1-PF	GTOT	Our	L1-PF	GTOT	Our
Hover	0.42	0.59	0.65	0.31	0.51	0.57
WalkingOcc1	0.36	0.23	0.52	0.42	0.23	0.45
WalkingOcc2	0.50	0.76	0.78	0.34	0.62	0.65
WalkingScale2	0.52	0.52	0.64	0.68	0.52	0.63
LightOcc	0.42	0.43	0.62	0.57	0.43	0.62
Shadow	0.62	0.63	0.52	0.60	0.53	0.58
FastWalk2	0.37	0.38	0.60	0.63	0.39	0.78
Walkingnight	0.45	0.49	0.53	0.53	0.47	0.66
Talking	0.37	0.66	0.60	0.49	0.64	0.57
Average	0.45	0.52	0.60	0.51	0.48	0.61

Finally, we randomly select two video pairs as example to illustrate the qualitative tracking performance of our method. From Fig. 4 we can clearly see that our method and GTOT can give a better tracking performance than other methods because they can effectively use the thermal information to enhance the robustness of visual tracking in grayscale video sequences.

6. CONCLUSION

In this paper, we have proposed an inverse sparse representation based tracking framework. This framework has benefited from the use of a correlation analysis based inverse sparse representation model that can jointly encode the target candidates in the graysecale and thermal video sequences. It has been shown through experiments that the proposed tracking framework can give a superior performance in challenging video sequences.

7. REFERENCES

[1] R. Gade and T. B. Moeslund, "Thermal cameras and applications: a survey," *Machine Vision and Applications*, vol. 25, no. 1, pp. 245–262, 2014.

[2] C. Li, H. Cheng, S. Hu, X. Liu, J. Tang, and L. Lin, "Learning collaborative sparse representation for grayscale-thermal tracking," *IEEE Transactions on Image Processing*, vol. 25, no. 12, pp. 5743–5756, 2016.

[3] X. Zhang, D. S. Pham, S. Venkatesh, W. Liu, and D. Phung, "Mixed-norm sparse representation for multi view face recognition," *Pattern Recognition*, vol. 48, no. 9, pp. 2935–2946, 2015.

[4] X. Mei and H. Ling, "Robust visual tracking and vehicle classification via sparse representation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 11, pp. 2259–72, 2011.

[5] C. Li, X. Sun, X. Wang, L. Zhang, and J. Tang, "Grayscale-thermal object tracking via multitask laplacian sparse representation," *IEEE Transactions on Systems Man and Cybernetics Systems*, vol. 47, no. 4, pp. 673–681, 2017.

[6] D. Wang, H. Lu, Z. Xiao, and M. H. Yang, "Inverse sparse tracker with a locally weighted distance metric," *IEEE Transactions on Image Processing*, vol. 24, no. 9, p. 2646, 2015.

[7] B. Zhuang, H. Lu, Z. Xiao, and D. Wang, "Visual tracking via discriminative sparse similarity map," *IEEE Transactions on Image Processing*, vol. 23, no. 4, pp. 1872–1881, 2014.

[8] S. Sun, X. Xie, and M. Yang, "Multiview uncorrelated discriminant analysis," *IEEE Transactions on Cybernetics*, vol. 46, no. 12, pp. 3272–3284, 2016.

[9] X. T. Yuan, X. Liu, and S. Yan, "Visual classification with multitask joint sparse representation," *IEEE Transactions on Image Processing*, vol. 21, no. 10, pp. 4349–4360, 2012.

[10] M. W. Schmidt, E. V. D. Berg, M. P. Friedlander, and K. P. Murphy, "Optimizing costly functions with simple constraints: A limited-memory projected quasi-newton algorithm," *Hansen Int*, vol. 5, no. 2-3, pp. 355–357, 2009.

[11] S. Zhang, X. Yu, Y. Sui, S. Zhao, and L. Zhang, "Object tracking with multi-view support vector machines," *IEEE Transactions on Multimedia*, vol. 17, no. 3, pp. 265–278, 2015.

[12] F. Liu, T. Zhou, C. Gong, K. Fu, L. Bai, and J. Yang, "Inverse nonnegative local coordinate factorization for visual tracking," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 28, no. 8, pp. 1752–1764, 2018.

[13] J. W. Davis and V. Sharma, "Background-subtraction using contour-based fusion of thermal and visible imagery," *Computer Vision and Image Understanding*, vol. 106, no. 2, pp. 162–182, 2007.

[14] M. H. Yang, H. Lu, and W. Zhong, "Robust object tracking via sparsity-based collaborative model," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2012, pp. 1838–1845.

[15] S. Hare, S. Golodetz, A. Saffari, V. Vineet, M. M. Cheng, S. Hicks, and P. Torr, "Struck: Structured output tracking with kernels," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 38, no. 10, pp. 2096–2109, 2015.

[16] M. Danelljan, F. S. Khan, M. Felsberg, and J. V. D. Weijer, "Adaptive color attributes for real-time visual tracking," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 1090–1097.

[17] Y. Wu, E. Blasch, G. Chen, L. Bai, and H. Ling, "Multiple source data fusion via sparse representation for robust visual tracking," in *International Conference on Information Fusion*, 2011, pp. 1–8.