Future Generation Computer Systems 113 (2020) 298–303

Contents lists available at ScienceDirect

Future Generation Computer Systems

journal homepage: www.elsevier.com/locate/fgcs

Learning multiple instance deep quality representation for robust object tracking

Guan Wang [a], Jing Liu [a,*], Wei Lo [a], Chun-Ming Yang [b]

[a] School of Business Administration, Guangxi University of Finance and Economics, No. 189, Daxuexi Road, Xixiangtang District, Nanning, Guangxi, 530007, China

[b] School of Economics and Management, Dongguan University of Technology, No. 1, Daxue Road, Songshan Lake, Dongguan, Guangdong, 523808, China

ARTICLE INFO

Article history:
Received 29 March 2020
Received in revised form 8 July 2020
Accepted 11 July 2020
Available online 13 July 2020

Keywords:
Visual object tracking
Quality model
Bidirectional LSTM
Weakly-supervised
Spatial temporal modeling

ABSTRACT

Robustly tracking various objects within a video stream with complex objects and backgrounds is a useful technique in next generation computer vision systems. However, in practice, it is difficult to design a successful video-based object tracking system due to the varied light conditions, possible occlusions, and fast-moving objects. In this work, a novel weakly-supervised and quality-guided visual object tracking model is proposed, wherein the key is a bidirectional long short-term memory recurrent neural network (BLSTM-RNN) that captures the feature sequence and predicts the quality score of each candidate window. More specifically, given a rich set of training videos annotated with the target objects, a weakly-supervised learning algorithm is first used to project all the candidate window features onto the semantic space. Next, we propose a two-stage algorithm to select the key frames from the video sequences, where both the shallow and deep filtering operations are conducted. Subsequently, the so-called BLSTM-RNN is proposed to characterize the feature sequence temporally, based on which the maximally possible object window can be calculated and finally output. In our experiment, a large video dataset containing 2045 NBA regular seasons and playoff basketball games was compiled. Based on this, a comparative study is conducted between the proposed algorithm and state-of-the-art video tracking methods. Extensive visualization results and comparative tracking precisions show the competitiveness of the proposed method.

1. Introduction

As a useful technique in computer vision, visual object tracking has attracted significant attraction in the past decades. Visual object tracking has many applications in daily life, such as abnormal event detection, intelligent automatic driving, and population flow analysis. For example, currently in China, there are many intelligent cameras installed in traffic intersection areas. In this way, unusual events such as pedestrians illegally crossing the road can be effectively tracked and flagged. As another example, in existing self-driving vehicle systems, it is necessary to integrate an effective and efficient object tracking system. When driving a vehicle from location A to B, a visual tracking model is needed to quickly identify possible obstacles, such as humans and guideposts, allowing the intelligent vehicle to reach the destination successfully. Moreover, in some crowded spots like the railway stations, it is important to inspect population flow by leveraging the camera network, so that abnormal population flow can be effectively detected and documented so that persons involved can be accurately located.

In the literature, tens and hundreds of visual object tracking models have been proposed and have achieved satisfactory performance. Some robust visual tracking algorithms have been commercialized and applied onto many domains, such as smart cities, human–computer interaction systems, and intelligent pedestrian inspecting systems. Although these visual tracking systems are successful, the current visual tracking technologies still suffer from the following shortcomings:

(1) To the best knowledge, there are many shallow/deep features characterizing the various objects within each candidate window. However, the importance of each kind of feature remains unknown. Based on the authors' experiences, the weights of different features are determined on a specific hand gesture extraction environment. For example, for robust object recognition under poor lighting conditions, more emphasis is placed on the silhouette feature while suppressing the rest. Meanwhile, when the target object motion is ultra-fast, a large weight can be assigned to the optical flow feature, while the other features can be

* Corresponding author.
E-mail address: liujing_ed@hotmail.com (J. Liu).

https://doi.org/10.1016/j.future.2020.07.024
0167-739X/

deemphasized. In practice, it is desirable to design a highly descriptive visual feature that can represent the target objects both temporally and spatially. Moreover, an optimal visual tracking system should exploit the inter-frame structure and successfully select the key video frames for processing.

(2) Existing visual tracking systems, are all trained by leveraging massive-scale training videos with manually annotated objects inside each video frame. This manual annotation process is expensive in practice. Usually, the human resources needed to annotate the target objects frame-by-frame for the. number of tracking videos are not available. Even worse, for each video frame, there can be multiple target objects with various scales. Therefore, there is a need to carefully identify each of these different-sized objects within a tracking video. This is a challenging task because of human errors. To reduce the labor-intensive human labeling process, it is necessary to leverage the weakly-supervised learning algorithms. However, designing such an algorithm remains a difficulty.

(3) As far as is known, the successful existing visual tracking models typically design each application-specific shallow feature. This strategy is effective and simple-to-implement, but relies heavily on the domain experiences of the system designer. That is to say, the visual tracking performance will be sub-optimal if the system designer is less qualified. In reality, a deep visual tracking framework that can automatically engineer visual features in a black-box way is needed. In this way, an experienced system designer to carefully develop the visual feature for tracking is required. However, designing an efficient deep model that optimally extract visual features for tracking is a difficult challenge. The specific problems include: (a) how to make the deep tracking model accurately encode the temporal-level visual features between video frames, and (b) how to enforce the learned deep model to incorporate the fine-grained spatial features of those tiny objects.

To solve, or at least alleviate, the previously mentioned limitations, this work proposes a novel, weakly-supervised, deep architecture for visual tracking, wherein only video-level semantic annotations are required. The flowchart of this proposed model is elaborated on in Fig. 1. In particular, given a rich number of training videos that are semantically labeled at the video-level, a novel two-stage video key frames extraction algorithm is developed in order to obtain the most representative video frames, wherein the two-stage operations abandon the less important video frames using shallow and deep visual features respectively. Afterward, a weakly-supervised learning paradigm to encode the video-level semantic tags into various video regions, by leveraging a manifold-based embedding algorithm is proposed. Afterward, a spatio-temporal convolutional neural network (SCN) is formulated to extract both spatial and temporal visual cues from the input video sequence. Lastly, a bidirectional long short-term memory recurrent neural network is formulated to model the feature sequence and calculate the quality score of each candidate window. Based on this quality score, the candidate window with the maximum quality score is output. Comprehensive experimental results and visualized tracking windows on the collected large-scale NBA playoff video dataset have demonstrated the usefulness of the proposed method in localizing various moving athletes.

The key contributions of the proposed weakly-supervised tracking model can be briefly described as follows: (1) A novel two-stage hybrid framework to extract the key video frames effectively is proposed. (2) A weakly-supervised learning framework to capably integrate video-level semantic tags into different video regions. (3) A spatio-temporal convolutional neural network (SCN) is formulated to extract both spatial and temporal visual cues from the input video sequence. and (4) A bidirectional long short-term memory recurrent neural network that describes the feature sequence and calculates the quality score of each candidate window is given.

The remainder of this article is organized in the following manner. In Section 2, previous works closely related to the work in this paper are briefly described and compared. In Section 3, a flowchart of the proposed weakly-supervised visual tracking models is given which details the three key contributions. Section 4 shows empirically how the proposed method performs and demonstrates its superiority. The last section reiterated the conclusions of the research and suggests some future work.

2. Related work

The proposed weakly-supervised visual object tracking model is closely related to research topics in computer vision and machine learning such as deep learning-based visual representation learning; and recurrent neural networks for visual modeling.

2.1. Deep neural network for visual feature extraction

Conventional image/video modeling algorithms leverage hand-crafted features such as SIFT (Scale Invariant Feature Transform) [1] and HOG (Histogram of Oriented Gradients) [2] for visually representing geometry and reconstructing it. In 2006, Hinton et al. [3] proposed the concept of hierarchical visual representation learning, and its layer-by-layer training algorithm can train deep neural networks effectively. After that, deep learning and neural networks once again gained the attention of computer vision researchers and are pervasively adopted in visual classification and retrieval, speech recognition, object recognition and other domains [4]. A few representative works are briefly described here.

Ji et al. [4] investigated RGB video-based human behavior recognition by formulating a three-dimensional convolutional neural network (CNN) model. They first used a series of fixed kernel functions to generate multi-channel information to characterize each video frame, and subsequently leveraged the 3-dimensional convolution to describe human motion information between multiple adjacent frames. Finally their work combines the information of all channels to obtain the final feature representation. Le et al. [5] combined the advantages of independent subspace analysis (ISA) and CNN. They first used the invariant spatiotemporal features in the ISA learning behavior video, and then incorporated the features as input to the multi-layer CNN network, thus utilizing CNN to learn higher-level and more abstract features of behavioral video. Du et al. [6] extracted the spatial-level and temporal-level deep features from the optical flow between video frames and consecutive frames, and subsequently used two deep networks to extract high-level features for human action recognition. Further in [7], Tran et al. proposed to model RGB domain videos as the research object and used the features extracted by 3-dimensional convolution directly so as to achieve a highly competitive recognition accuracy.

2.2. Recurrent neural network for computer vision applications

Recurrent neural networks (RNNs) are highly competitive techniques used to extract hidden visual features in a temporal or spatial domain sequence. The history of recurrent neural networks dates back to the Elman RNN in early 1990s [8]. Although the RNN was originally designed for modeling the long-term dependencies, a large number of applications have demonstrated

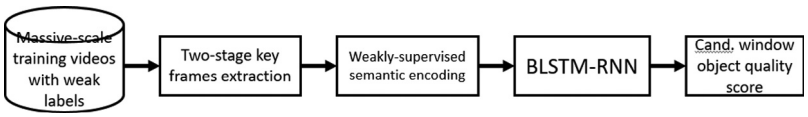

Fig. 1. The pipeline of the proposed weakly-supervised visual tracking framework.

that standard RNNs often have difficulties in achieving long-term preservation of visual information. Bengio et al. [9] proposed that the standard RNN has the problem of gradient disappearance and gradient explosion. Both of these problems are caused by the iterative nature of RNN, and therefore RNN could not be pervasively applied in the early days. To solve the problem of long-term dependence, Hochreiter et al. [10] proposed a Long-Short Time Memory (LSTM) network to upgrade the traditional RNN model. LSTM has also become the most efficient sequence model in current practical applications. Compared to the hidden unit of RNN, the internal structure of the hidden unit of LSTM is highly complex. By incorporating linear intervention during the flow of information along the deep network, LSTM can selectively add or reduce visual features. RNN has a variety of excellent inherent structures, such as the Gated Recurrent Unit (GRU), which is popularly adopted in practice. Both LSTM and GRU maintain long-term dependencies by adding internal gating mechanisms. Their loop architecture only has dependencies on the entire set of past states. Accordingly, the current state may also depend on the future state. Schuster et al. [11] proposed a bidirectional neural network (BRNN) which can learn the context in two directions in the temporal domain. BRNN involves two different hidden layers, and the input is obtained in two directions. Graves et al. [12] used bidirectional LSTM (BLSTM) to achieve excellent performance in phoneme recognition.

3. Our proposed method

3.1. Two-stage key frame selection

In our weakly-supervised visual tracking framework, a two-stage scheme is first used to extract the key frames from a set of videos. The filtering of video image key frame sequences is vital to high accuracy and real-time performance of video feature extraction. If all the frames in the video image sequence are used, not only will it take a lot of time, but the performance of following procedures will be degraded. If the baseline between two adjacent frames of the video image sequence is too small, the accuracy of triangulation will be reduced, and if it is too large it will increase the difficulty of feature matching. A novel two-step algorithm for video frame sequence filtering is designed. Multi-layer theory is introduced to divide key frames into shallow and deep layers for filtering. The shallow filtering step considers the image clarity evaluation. After the key frames are filtered in the shallow layers, factors such as the overlap degree of the key frames and the width of the baseline, which have a huge influence on the deep filtering, are considered.

Image degradation occurs when the key sequence of the video images is segmented and the camera unexpectedly moves vigorously due to environmental factors. Therefore, shallow filtering mainly considers these two factors. However, in the process of key frame filtering, there is no example or template as a reference, thus the problem can be classified as a no-reference image quality assessment problem. In this case typical quality assessment criteria such as sharpness and gradient functions of the image are considered:

(1) Brenner gradient. The Brenner gradient function calculates the squared difference of the gray values of adjacent pixels to complete the resolution evaluation:

$$D(f) = \sum_y \sum_x |f(x+2, y) - f(x, y)|^2, \tag{1}$$

where $f(x, y)$ is the gray value of position (x, y), and $D(f)$ is the evaluation result.

(2) Tenengrad gradient. The Tenengrad gradient is defined by:

$$D(f) = \sum_y \sum_x |G(x, y)|, G(x, y) > T, \tag{2}$$

and

$$G(x, y) = \sqrt{G_x^2(x, y) + G_y^2(x, y)} \tag{3}$$

where T is the threshold for edge detection, G_x is the convolution of the edge detection operator at the pixel (x, y) in the horizontal direction by the Sobel operator, and G_y is the convolution of the edge detection operator at the pixel (x, y) in the vertical direction by the Sobel operator. The template is:

$$g_x = \frac{1}{4}\begin{bmatrix} -1 & 0 & 1 \\ -2 & 0 & 2 \\ -1 & 0 & 1 \end{bmatrix}, g_y = \frac{1}{4}\begin{bmatrix} 1 & 2 & 1 \\ 0 & 0 & 0 \\ -1 & -2 & -1 \end{bmatrix}. \tag{4}$$

(3) Laplacian gradient. The Laplacian operator is defined as:

$$\mathrm{L} = \frac{1}{6}\begin{bmatrix} 1 & 4 & 1 \\ 4 & -20 & 4 \\ 1 & 4 & 1 \end{bmatrix}, \tag{5}$$

And the definition for image quality is:

$$D(f) = \sum_y \sum_x |G(x, y)|, G(x, y) > T \tag{6}$$

where $G(x, y)$ is the convolution of Laplacian at (x, y).

The shallow filtering step mainly filters the maps of all the frames of the video according to the evaluation function. After the calculation, a threshold σ is obtained. After normalization using this threshold σ, a peak is generated in a similar time interval, and the peak represents the highest resolution of the video image of the frame. The frame image is output as a result of shallow filtering. After shallow filtering, it is important to continue to improve the efficiency of modeling, and to avoid the problem of triangulation accuracy degradation and motion degradation due to the high similarity of the two images. Geometric robust information criterion (GRIC) and feature matching degree are used as conditions for deep filtering of the results of shallow filtering. The basic matrix will sometimes cause large calculation errors, due to degradation phenomena such as motion degradation and structural degradation between the two key frame images. The basic matrix is largely considered as the basis of camera estimation and video quality assessment. Therefore, GRIC is used to distinguish between sports degradation and other phenomena. GRIC is defined as:

$$\mathrm{GRIC} = \sum_i \rho\left(e_i^2\right)_i + \lambda_1 dn + \lambda_2 k \tag{7}$$

$$\rho\left(e_i^2\right)_i = \min\left(\frac{e_i^2}{\sigma^2}, \lambda_3(r-d)\right) \tag{8}$$

where e_i is the number of residues, $\lambda_1 = \ln r$, $\lambda_2 = \ln(rn)$, λ_3 is the value that controls the residue number, N is the number of matching points of the two images, K is the model freedom degree, ($k = 7$ for the homograph matrix model and $k = 8$ for the feature matrix model), σ^2 is the variance of the error, r is

the dimension of the data, and d is the dimension of the model. For the key frame of the video image to be matched, the value of the base matrix F and the value of the GRIC of the homograph matrix H are respectively calculated. If GRICF < GRICH, then it can be stated that the two frame images do not have a degenerate relationship, and the frame is listed as a candidate key frame.

3.2. Weakly-supervised semantic encoding

Each semantic video-level labeled video contains a bag of candidate windows. It is natural and straightforward to enforce the following three constraints between the semantic label at image-level and the quality score of each candidate window, that is: (1) Each candidate window should correspond to at most one quality score.; (2) Each video-level quality score has at least one candidate window with the equivalent quality score.; and (3) Quality scores of candidate windows from each video should be maximally consistent with the ground-truth quality score at video-level. Based on these three constraints, the following objective function is minimized:

$$Q\left(\boldsymbol{R}\right)=\sum_{i=1}^{v}\sum_{j=1}^{N_t}\left|\max_{r\in J_i} r_{ij}^{c}-g_i^{c}\right|=,\qquad \mathbf{R}^{\mathrm{T}}\mathbf{R}=\mathbf{I}. \tag{9}$$

After some derivations, the above formulation can be re-organized into:

$$\mathrm{Q}\left(\mathbf{R}\right)=\sum_{i=1}^{v}\sum_{j=1}^{N_t}[\left(1-g_i^{c}\right)l_c^{T}\boldsymbol{R}^{T}+g_i^{c}(1-\max_{r\in J_i}\phi_{ij}^{T}\boldsymbol{R}l_c)], \tag{10}$$

By combining the above two optimization functions, the final objective function for the ranking algorithm is:

$$\min_{\mathrm{R},\mathbf{C}}\sum_{i=1}^{v}\sum_{j=1}^{N_t}\left[\left(1-g_i^{c}\right)l_c^{T}\boldsymbol{R}^{T}+g_i^{c}\left(1-\max_{r\in J_i}\phi_{ij}^{T}\boldsymbol{R}l_c\right)\right],$$
$$\text{s.t.},\ \mathbf{R}^{\mathrm{T}}\mathbf{R}=\mathbf{I}, \tag{11}$$

Noticeably, it is infeasible to solve (11) analytically. In this implementation, the well-known concave convex procedure (CCCP) [12,13] is used to solve it iteratively.

3.3. BLSTM for sequence modeling

We use the BLSTM to model the generated feature sequence and output the final quality score. Given a input sequence $\mathrm{x}=(x_1, x_2, \ldots, x_T)$, a standard RNN iteratively calculates the state sequence $\mathrm{h}=(h_1, h_2, \ldots, h_T)$ and output sequence $\mathrm{y}=(y_1, y_2, \ldots, y_T)$, which can be detailed as follows:

$$h_t=f(W\left[x_t, h_{t-1}\right]+b_h) \tag{12}$$

$$y_t=W_{hy}h_t+b_y \tag{13}$$

In (13) and (14), $\mathbf{W}$ represents the weight matrices, b_h and b_y are bias terms in the hidden and output layers, respectively, and f is an activation function in the hidden layer. The traditional RNN does not work well for long-sequence feature modeling due to the gradient vanishing problem. The long short-term memory (LSTM) solves the problem of gradient vanishing by constructing a memory unit to model the sequence information [10,14]. The traditional RNN can only record the forward context information, but the features of the current frame are highly related to adjacent frames in both the forward and backward directions. The bidirectional recurrent neural network is capable of recording bidirectional context information before and after the feature sequence [11,15], which divides the hidden layer into a forward sequence and a backward sequence.

In the process of modeling the video feature sequence based on the BLSTM, model training is carried out by minimizing the model output the error squared sum (SSE) between the estimated and the real quality score. For the input sequence l with the frame number Ml, the error function is defined as

$$E_l\left(w\right)=\frac{1}{2}\sum_{m=1}^{M_l}E_{lm}\left\|o_m^{l}-\hat{o}_m^{l}\right\|^2 \tag{14}$$

In iteration t, the error gradient is

$$\Delta w\left(t\right)=m\Delta w\left(t-1\right)-\alpha\frac{\partial E(w\left(t\right))}{\partial w(t)}, \tag{15}$$

where α is the learning rate of the RNN model, and m is a momentum parameter. The image quality score is obtained by iteratively calculating (14) and (15) based on the training dataset.

Based on the learned BLSTM model, a Gaussian mixture model is used to characterize each annotated target object:

$$prob\left(f\right)=\sum_{i=1}^{H}\tau_i\cdot prob(f|\theta_{\mathrm{i}}, \Theta_{\mathrm{i}}), \tag{16}$$

where f denotes the deep feature learned using the BLSTM model for the candidate window, H is the number of the Gaussian components, τ_{i} denotes the weight that indicates the importance of the ith Gaussian component, and θ_{i} and Θ_{i} are respectively the mean and variance of the ith Gaussian component.

4. Experimental results and analysis

In order to test the video quality assessment algorithm proposed in this paper, a total of 645 videos in which the video frame rate is 30 fps from the NBA regular season and playoff basketball games from 2016 to 2018 were used as the original training data.

To reduce the impact on different experimental results, each video is pre-processed before the experiment. (1) Background removal: In depth video, the depth information of the background is consistent, and the depth information of the foreground is variable. Therefore, the background information can be removed for each feature. (2) Bounding box determination: For each video, according to each frame, a bounding box is obtained that frames the human behavior, and the maximum of all frames is obtained. The bounding box is used as the bounding box of that particular video. (3) Normalization: All of the videos processed in the previous step are normalized to a uniform size using an interpolation technique, where the normalized number of video frames is equal to the middle of all of the video frames. At the same time, the depth information value of all videos is normalized to the range [0,1] using the min–max method. Finally, all samples are horizontally flipped to form a new sample, thereby multiplying the training samples in the dataset. The experimental deep neural network model was written using the Caffe platform [16,17], and the data pre-processing part was completed using Matlab.

4.1. Comparative study

In this subsection, the weakly-supervised visual tracking using our complied NBA playoff video datasets. We spent lots of human resources to semantically annotate the players' name, the team name, and location of each NBA video game. The dataset was divided evenly into two disjoint sets, one used for tracking model training and the other used for the tracking model for model evaluation. The experiment was carefully designed to ensure that no trajectories or targets that appear in the evaluation dataset. For the entire set of NBA playoff video sequences, only the two-dimensional xy-coordinate videos without the location information is experimented. Moreover, the videos are

Table 1
Comparative average visual tracking performance.

	Breit	Kuo	Huang	Wu	Angeivoa	Proposed
MOTP	0.5654	0.6543	0.5860	0.6675	0.7121	0.7765
MA	0.1121	0.1010	0.1322	0.1232	0.0921	0.0632
FP	2%	2.2%	2.1%	2.4%	2.3%	2.0%
IDS	19	12	14	11	13	10

not pre-processed with calibration and entry/exit zones. All of the parameters with the exception of the inherent deep model parameters, via cross validation. The weakly-supervised visual tracking is compared with a series of the state-of-the-art visual object tracking algorithms.

To understand the comparative results relative to the different visual tracking models, the well-known CLEAR MOT metric [17,18] is adopted for model analysis. This metric can output a precision score for each compared model; the multi-objective tracking precision (MOTP), and an object tracking quality score; the multi-object detection accuracy (MOTA) [19,20]. We also report the missing rate (MR), the ratio of false positives (FP), and the identity switches (IDS). During the comparison, a continuous projection between the pre-assumed tracking position and the ground-truth in each video frame. Based on these observations, each of the aforementioned tracking performance evaluation metrics is computed in the following manner[19,21]. For each pair of objects, the tracking precision score is defined using the intersection between unions of their bounding boxes. In addition, the MOTP is calculated as the score averaged by the entire number of matched pairs. Based on experience, a larger MOTP value indicates a higher accuracy. The MOTA denotes the percentage of the correct detections [22]. A tracking failure happens if an NBA player is mistakenly annotated, but the intersection between unions is less than 0.5. The comparative visual tracking performance is shown in Table 1. As can be seen, the proposed method received the highest tracking accuracy.

4.2. Important parameter analysis

The only tuning parameter in the proposed weakly-supervised visual tracking is the number of Gaussian components, H. In this experiment, the MOTP value of our weakly-supervised visual tracking model is reported. Herein, both the single object-based visual tracking and multiple objects-based visual tracking are evaluated [13,23]. As shown in Fig. 2, the proposed method performs the best when H = 4, for both single object and multiple object tracking.

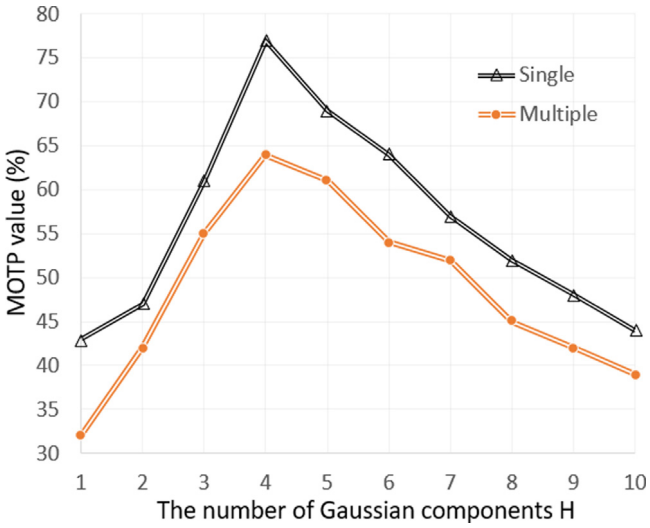

Fig. 2. Visual object tracking performance by varying H.

5. Conclusions

In this work, novel weakly-supervised video tracking model is proposed to effectively detect objects from massive-scale videos. Given a rich set of videos collected a weakly-supervised learning paradigm was formulated to map the entire candidate windows onto a manifold-based semantic space. Next, a two-stage algorithm to select the key frames from the video sequences was proposed. During the selection process, both the shallow and deep filtering operations were carried out. Finally, designed the BLSTM-RNN framework was designed to represent the feature sequence in the temporal domain. Afterward, the maximally possible object window is output as the visual tracking results. Comprehensive experimental results have fully validated the usefulness and robustness of our method.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was financially supported by National Social Science Foundation of China, under grant number 19BJY066; Social Science Foundation of Ministry of Education of China, under grant number 18YJC790159

References

[1] David G. Lowe, Object recognition from local scale-invariant features, iccv 99 (2) (1999).
[2] Navneet Dalal, Bill Triggs, Histograms of oriented gradients for human detection, in: International Conference on Computer Vision & Pattern Recognition (CVPR'05). Vol. 1, IEEE Computer Society, 2005.
[3] Geoffrey E. Hinton, Simon Osindero, Yee-Whye Teh, A fast learning algorithm for deep belief nets, Neural Comput. 18 (7) (2006) 1527–1554.
[4] Shuiwang Ji, Wei Xu, Ming Yang, Kai Yu, 3D Convolutional neural networks for human action recognition, IEEE Trans. Pattern Anal. Mach. Intell. 35 (1) (2012) 221–231.
[5] Quoc V. Le, Zou Will, Serena Yeung, Andrew Y. Ng, Learning hierarchical invariant spatio-temporal features for action recognition with independent subspace analysis, 2011.
[6] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, Manohar Paluri, Learning spatiotemporal features with 3d convolutional networks, in: Proceedings of the IEEE international conference on computer vision, 2015, pp. 4489-4497.
[7] Karen Simonyan, Andrew Zisserman, Two-stream convolutional networks for action recognition in videos, Adv. Neural Inf. Process. Syst. (2014).
[8] Jeffrey L. Elman, Finding structure in time, Cogn. Sci. 14 (2) (1990) 179–211.
[9] Bengio Yoshua, Patrice Simard, Paolo Frasconi, Learning long-term dependencies with gradient descent is difficult, IEEE Trans. Neural Netw. 5 (2) (1994) 157–166.
[10] Sepp Hochreiter, Jürgen Schmidhuber, Long short-term memory, Neural Comput. 9 (8) (1997) 1735–1780.
[11] Mike Schuster, Kuldip K. Paliwal, Bidirectional recurrent neural networks, IEEE Trans. Signal Process. 45 (11) (1997) 2673–2681.
[12] Darya Ismailova, Wu-Sheng Lu, Penalty convex-concave procedure for source localization problem, in: CCECE, 2016, pp. 1–4.
[13] Mario Frank, Andreas P. Streich, David Basin, Joachim M. Buhmann, Multi-assignment clustering for boolean data, JMLR 13 (2012) 459–489.
[14] Tetsuya Yoshida, Toward finding hidden communities based on user profile, in: ICDMW, 2010.
[15] Jure Leskovec, Kevin J. Lang, Michael W. Mahoney, Empirical comparison of algorithms for network community detection, in: Proc. Of WWW, 2010.

[16] Yangqing Jia, Evan Shelhamer, Jeff Donahue, Sergey Karayev, Jonathan Long, Ross Girshick, Sergio Guadarrama, Trevor Darrell, Convolutional architecture for fast feature embedding, in: Proceedings of the 22nd ACM International Conference on Multimedia, ACM, 2014, pp. 675–678.
[17] Andrea Lancichinetti, Santo Fortunato, Filippo Radicchi, Benchmark graphs for testing community detection algorithms, Phys. Rev. E 78 (2008) 046110.
[18] K. Bernardin, R. Stiefelhagen, Evaluating multiple object tracking performance: The CLEAR MOT metrics, J. Image Video Process. 2008 (2008) 1–10.
[19] Steve Gregory, A Fast Algorithm to Find Overlapping Communities In-Networks, Machine Learning and Knowledge Discovery in Databases, Vol. 5211, Springer Berlin Heidelberg, 2008, pp. 408–423.
[20] Jaewon Yang, Jure Leskovec, Community affiliation graph model for overlapping network community detection, in: ICDM, 2012.
[21] Gergely Palla, Imre Dernyi, Ills Farkas, Tams Vicsek, Uncoveringthe overlapping community structure of complex networks in nature and society, Nature 435 (7043) (2005) 814–818.
[22] Yuzhou Zhang, Jianyong Wang, Yi Wang, Lizhu Zhou, Parallel community detection on large networks with propinquity dynamics, in: SIGKDD, 2009.
[23] Tianbao Yang, Rong Jin, Yun Chi, Shenghuo Zhu, Combining link and content for community detection: A discriminative approach, in: ACM SIGKDD, 2009.

Guan Wang received the M.S. degree in business administration from School of Economics and Management, Hennan Polytechnic University, Hennan, China, in 2007, and the Ph.D. degree in Mining Engineering from School of Business Administration, Hennan Polytechnic University, Hennan, China, in 2016. In 2017, he joins the Guangxi University of Finance and Economics, China, where he is currently an Associate Researcher with the School of Business Administration. His research interests mainly include resource development and regional sustainable development.

Jing Liu received the Master of Science in Management from School of Economics and Management, Hennan Polytechnic University, Hennan, China, in 2011. In 2018, she joins the Guangxi University of Finance and Economics, China, where she is currently an Associate Professor with the Graduate Institute. Her research interests mainly include business administration and regional economics.

Wei Lo received the master degree of business administration from National Yunlin University of Science and Technology, Yunlin, Taiwan, in 2003, and the Ph.D. degree in business administration from Fu Jen Catholic University, New Taipei City, Taiwan, in 2018. In 2018, he joins the Guangxi University of Finance and Economics, China, where he is currently an Associate Professor with the School of Business Administration. His research interests mainly include entrepreneurship and innovation.

Chun-Ming Yang received the M.S. degree in business administration from Asia University, Taichung city, Taiwan, in 2010, and the Ph.D. degree in management sciences from Tamkang University, New Taipei City, Taiwan, in 2015. He is currently an Associate Professor at the School of Economics and Management in Dongguan University of Technology, China and his research interests mainly include process capability analysis, quality management, and decision-making.