Long-term tracking of human skin areas on thermal images for monitoring vital signs

Seunghoi Kim

16060231

Abstract

This paper reviews state of the art of the methods for tracking human skin tissues on thermal images and evaluates the approaches for the use in the applications for monitoring human vital signs Many tracking methods from classical template matching to recent machine learning techniques are reviewed, compared and evaluated in terms of how they will fit to long term tracking of human tissues on thermal images. On thermal images having lower resolution, more blurred edges and less texture over visible images, the tracking require to be fast, accurate and robust enough to monitor vital signs faithfully for a long term. Although some conventional computer vision based methods like KCF have shown top performance so far, many recent CNN-based methods have been gaining popularity and outperforming conventional computer vision methods and machine learning methods showing a good performance in recent years.

1 Introduction

Although the researches on tracking in thermal images have not done as much as RGB, the importance of accurate tracking on thermal image has been rising due to increasing demands in HCI, medical and psychological fields. Thermal images provide a vital data unobtrusively that visible images cannot provide. By detecting vital signs from thermal image that sensors in visible domain cannot detect [5], for example, respiratory rate and heart rate can be measured [5]. Furthermore, thermal imagesare not affected by light illumination unlike visible images [2], , they have an advantage over visible images in monitoring vital signs in the fields of medicine and psychology. However, because thermal images have lower contrast less skin texture, different approaches over methods developed on visible domain may be required [7]. It is crucial to have high reliability in tracking over a long period that means the tracker must be computationally light to achieve real time performance (at least 20 fps [17]). As one of the classical methods, Matthews and Ishikawa [1] introduced a template matching method, in which template is updated every frame using warp parameters to handle some extends of deformation. The problem in this method is that small error accumulates as template is updated, leading to 'drift problem'. Furthermore, the high computational cost limits its usage in real time applications. This review overviews various highlighting methods for tracking deformable objectsand evaluates the suitability of the methods in tracking human skin tissues in thermal images. Furthermore, challenges leading to another new research opportunities are discussed to further advance tracking performance in thermal images. The rest of this paper is organized as follows: Sections 2 to 7 introduce various tracking methods already published and Section 8 states major challenges and opportunities in future research. Finally, Section 9 addresses conclusion.

2 STM particle filter

For reliable tracking, it is essential to overcome challenges such as positional and physiological changes. For example, in sleep studies, if the tracker fails tracking the nasal ROI, the result of breathing signal will be inaccurate which affects accurate analysis [5]. For this reason, tracking using shape models by Cootes et al.[3] is not ideal because nose is normally colder than the surrounding which gives a characteristic thermal shape, but when there is a sudden physiological change like allergic reaction, the nose temperature will change significantly which neutralizes a shape model tracker. Hence Zhou et al.[5] proposes a STM particle filter which is specifically designed for tracking a facial tissue of interest reliably. They use particle filtering which approximates the posterior distribution to predict where the next ROI's position will be by giving weights to each particle. It uses 'maximum a posteriori'(MAP) to determine the state of the tracker by selecting the current ROI from the candidates of ROI. When the current ROI is selected based on a MAP estimate, the template is updated based on the formation of a Matte by Carsten Rother et al.[13](developed for segmentation purposes) from stable and unstable seeds, the previous template and the MAP ROI estimate. Then the updated template is used in the next time step. The method to extract stable and unstable seeds follow the Matte computation step which if $|F_t[i] - T_(t-1)[i]| < 1$ where t is time, then the pixel is stable and unstable if $|F_t[i] - T_(t-1)[i]| > 2$ [5] where lambda is predetermined thresholds. The Matte values are set to 0 or 1 at the locations of extremely unstable or stable pixels. Then, the computation of STM provides values of pixels in [0, 1] which indicates the stability probability of the corresponding pixel in the template. And these estimated Matte values (the weight) indicate the necessary degree of updating of each pixel. Finally, the template is updated using the equation 1,

$$T_t[i] = \alpha_t[i]T_t - 1[i] + (1 - \alpha_t[i])F_t[i]. \tag{1}$$

If $\alpha_{t[i]}$, the Matte value, is stable seed then the template value will stay same as the value will be 1. On the other hand, for the unstable seed, the template will be updated by the corresponding pixel in the current ROI, $F_{t[i]}$. Hence, this updated template is representation of the latest version of the object's appearance while it also reserves the stable pixels of the previous template to prevent drifting. The experiments show that STM particle filter outperforms OAM method and Zero- One method. Furthermore, STM shows to track reliably during exhibition of large changes in position and physiology. The use of particle filter decreases chances of the tracker falling into local minima. Many tracking techniques track by gradient descent method, or mean shift which is effective but also can fall into local minima easily, if the subject abruptly changes its position in the image. However, by using a particle filter, it can predict where the subject will be in the next frame, so there is a less change of falling into local minima. Furthermore, this algorithm only shows the tracking of exclusive salient regions like nose which are easy to find due to their temperature but also only allows a limited robustness when there is a partial occlusion and out-of-plane rotation [5]. Hence, it will not be suitable

for tracking human skins especially when the human body such as nasal part exhibits a significant change in physiology affecting the temperature.

3 Active Appearance Model

All the template tracking methods take a high computational power, making it difficult to use in real time applications. Also, many methods listed above are done under a highly restricted environment such as limiting head movement for achieving a high accuracy. The method proposed by Kopaczka et al.[7], introduce a LWIR face tracker based on an Active Appearance Model which can track whole face robustly. Compare to other AAM, this approach extends and has unique steps: pre-processing images and AAM fitting. As shown in [8], the contrast-enhancing preprocessing such as antistrophic diffusion of LWIR images have a positive impact on the fitting performance of an AAM. Here, it implements sharpening filters. The methodology is that images are firstly filtered using preprocessing filter to enhance image contrast which makes it easier to extract features. Then, AAMs are trained with dataset such as images of persons with landmarks for facial regions like nose. The main components of appearance are extracted by using PCA (Principal Components Analysis) which is a feature extraction by performing dimensionality reduction and finding axis with the biggest variance. For AAM fitting, they use extension of Inverse-Compositional (IC) algorithms, simultaneous inverse compositional (SIC) and alternating inverse compositional (AIC) which simplifies iterative computation, leading to an increase in computation speed while keeping the fitting precision [8]. The results show that introduction of preprocessing algorithms improved the AAM fitting performance for intensity based AAM especially when fitting algorithms such as AIC or SIC is used but had small impacts on feature based AAM. Furthermore, unlike other methods explained above, this model performed well in tracking under conditions such as pose changes, fast head movement and extreme out-of-plane rotation. The experiment was also done in 30 frames per second, which is viable for the use in real-time applications, but the test was only done in 60 seconds of video sequence, so there is a question of suitability in long term tracking application.

4 Multi-Domain Convolutional Neural Networks

This approach (MDNet) by Nam et al [14] uses CNN which is pre-trained by using datasets including ground truth to acquire the tracking object information. The methodology is as follow: First, a 107*107 RGB input is received and trained by Stochastic Gradient Descent method which is a method to find a optimizing solution. Unlike classical Gradient Descent, it trains mini-bath by sampling data randomly first rather than a whole batch. Therefore, the learning rate is slower and it tends to have more noise but can escape local minimum easier than gradient descent. Hence, it updates the weight of nodes more frequently and finds the optimizing solution faster. Finally, the useful features

or information are obtained. They use 'Hard negative mining' approach to be used for real time by changing training and testing procedures to identify the false positives. The network only comprised with five hidden layers which includes three convolutional layers and two fully connected layers to reduce time cost[14]. However, due to this feature, it sometimes tend to fail to acquire the target tightly. The bounding box regression technique is hence introduced to improve the accuracy [14]. The MDNet is tested under VOT2014 and Objet Tracking Benchmark. The results show that it outperforms many other state of the art trackers and takes first place in accuracy and robustness in VOT2014. The algorithm also handles well under conditions of change in illumination, blurring and deformation. However, the fact that it runs at only 1fps show that it is very slow for real time application [14].

5 Tracking-Learning-Detection and median flow

None of the methods have yet shown that they are able to perform tracking images reliably and robustly. Another state-of-the-art technology, TLD by Kalal et al.[8], which is a semi-supervised learning, using PN tracker can track while learning. The tracker used in TLD is based on median flow, which is proposed by Kala et al. [4], a method of tracking and evaluating tracking failure. The main principle behind 'median flow' is that it measures the difference between the two trajectories performed by forward and backward (FB) tracking. Hence, with a combination of a point tracking algorithm, Lucas-Kanade tracker and FB error to evaluate errors develop a robust object tracker. FB error is calculated by evaluating forward and backward trajectory which is done by applying the same tracking on a reversed sequence of images. The reason behind the use of FB as an error measure rather than sum-of-square differences (SSD) is that SSD cannot detect inliners (displacements of the points between the original and warped images closer than 2 pixels from the ground truth) when a threshold for deciding inliner is small [4]. Hence, the tracker estimates displacements of a number of points within the object's bounding box and 50 percent of the worst predictions are filtered out while the remaining predictions are used to estimate the displacement of the whole bounding box. The novelty of this tracking method, TLD, is that it uses a tracker and a detector together. The tracker provides training data for the detector while detector detects an object based on the trained data. Hence, tracker can fail if it misses tracking the object even at one frame, but detector prevents this problem and drift by re-initializing the tracker. It also has PN learning which finds false negatives and false positives to estimate errors and retrain the detector to avoid the errors. The procedure of TLD is as follow: first the user picks out the region which initializes tracker and detector. Detector is trained with the first input image in the beginning and both tracker and detector are used to find the object from next frame. If tracker successfully finds the object, the window that tracker found is used as the data for training detector. Also, negative examples are extracted and trained to detector if the

region detector found is not same as the tracker. However, if tracker fails which means the median $|d_i$ - $d_m| > 10$ pixels, the tracker does not return the bounding box and allows detector do the job and reinitialize the tracker to initiate the tracking. Another novelty of this solution is that TLD uses an online detector which means it can perform in real time. The results show that TLD showed a higher performance comparing to other tracking algorithms such as Iterative Visual Tracking(IVT) and Co-trained Generative-Discriminative tracking (CoGD) and only CoGD performed similarly, while it only ran at 2 fps, TLD performed at 20 fps. However, tracking on objects such as pedestrian and a car chase, TLD performed less due to the limitation of it tracking articulated objects. Furthermore, TLD uses Median flow which I explained in the above which tracker does not perform well in the condition of full out of plane and tracking articulated objects which are the problems which 'median flow' carry as it will drift away from the target. Furthermore, TLD only trains the detector while tracker is fixed which means that the tracker will always make the same errors.

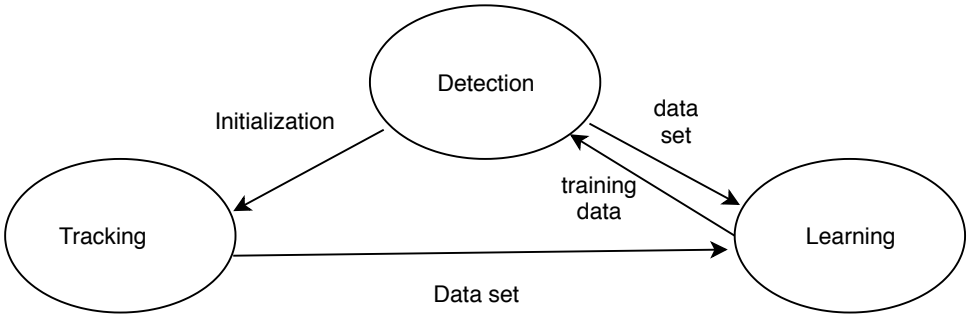

Figure 1: TLD

6 Kernelized Correlation Filters

Most modern trackers are discriminative classifier which has a problem with datasets that are redundant due to reasons like overlapping pixels. The new tracking method by Henriques et al.[9] is able to diagonalize the resulting data matrix with the discrete Fourier Transform and map nonlinear filter as fast as linear correlation filters by using kernel trick which reduces both storage and computational power. The principle of this algorithm is as follows: linear regression which aims to find $f(z) = w^T z$ that minimizes the squared error over samples x_i and their regression targets y_i [10]. Cyclic shifts to obtain all permutations of matrix given to train a classifier with a positive samples and translations of it. Circulant matrices to compute a regression with shifted examples. The circulant matrix is generated by the first row. Then the base sample is shifted by one in the next row and so on to form cyclic shifts of vector image. All circulant matrices are made diagonal by the discrete Fourier transform. This novel approach outperforms Struck and TLD and also runs faster at 100FPS which is very ideal for real time application. Furthermore, running the method on HOG features instead of raw pixels

increase the performance and the speed much higher. This algorithm is also tested in thermal imaging by Stojanovic et al.[10]. The paper uses KCF to test using LTIR dataset and it shows that it performs well but still struggles and lose tracking in the conditions such as overlapping, similar to background and very low contrast [10].

7 Edge boxes

This new method of tracking called, Edge Box Tracker (EBT) by Zhu et al.[14] uses a novel approach which probes entire frame to track an object rather than finding an object within the local window. Most trackers use local window rather than search in a whole frame due to high computational power, but EBT implemented it very efficiently. The methodology is as follows: The initial object location is manually put in first frame. From next frame, the number of candidates for bounding boxes are obtained from an image. Since the numbers are large, it must be computed efficiently to reduce the computational power. The edge map is achieved using Structured Edge detector [16]. Then the neighbouring edge pixels with similar orientation are grouped together and the affinities of edge groups are calculated. Finally, the score for an each box is calculated by adding up the edge strength of edge groups inside the box, minus the strength of edge groups that are part of a edge or contour of the box which then extracts the bounding box with the highest scored candidate. On top of that, it uses 'Instance Specific Proposals' which takes object instance level into an account to extract the top candidate for reliable tracking[15]. Incorporating with Structure Support Vector Machine, the bounding boxes that are higher than the predetermined thresholds are accepted to the classifier for re-ranking to reduce time cost. The updating of classifier is performed at every 5 frames. The test was performed under VOT protocol and it showed that EBT outperforms KCF. However, this excellent robustness of EBT is achieved due to enlarging a bounding box in cases of low tracking confidence [17]. In conditions such as fast moving objects and ultra-low-frame-rate (sampling sequence at every 20 frames), EBT shows a better performance than other top-ranked trackers such as KCF and MEEM [15]. Although, KCF remained the fastest in terms of computational power, as EBT (4.4fps) goes through a whole image, it does not fall behind compare to other ranked trackers. However, in order to be used in real time applications, it requires to be performed at much higher speed.

8 Challenges and research opportunities

Each tracking method from classical template updating method to the other machine learning methods reviewed in above sections shows pros and cons as summarized in the Table below.

Tracking method	Adavantage	Disadvantage
Template update	-Easy to implement	-Template drifting as errors accumulate during the update
AAM	-Endure significant plane rotations and fast head movements	-Cannot handle cases well outside of the training set such as occlusions and extremely deformable objects
Particle filter	-Handle nonlinear motions in the predict-update loop	-Low performance in fast movement and significant rotation
Multi-domain CNN	-Handle deformations and illumination changes	-High Computational cost
TLD	-Robust and relatively fast	-Low performance in the case of occlusion and disappearance of object
KCF	-Fastest out of all methods reviewed	-Lower performance in IR domain
EBT	-Highest performance in VOT-TIR 2016 challenge	-Slower than KCF and bounding box is enlarged during low tracking confidence

Over Template Matching as one of the most classical methods, the other machine learning methods reviewed here showed better performance in terms of speed, accuracy and robustness because they can make a good use of object statistics obtained from the training set. Although KCF showed the fastest performance among all the methods reviewed, it showed lower performance in thermal images featuring lower resolution, blurred edges and less texture. Multi-domain CNN method outperformed many other state of the art trackers and takes first place in accuracy and robustness in VOT2014. Furthermore it could also handle well change in illumination, blurring and deformation. However, it was too slow to use in real time. The current trend of the tracking method is using CNN features such as Multi-domain CNN. In VOT-2015 challenge, only two CNN-based features were listed while in VOT-2019, 69 percents used CNN features. This suggests the increase in popularity of CNNs as feature extractor. According to recent VOT challenges, many CNN- based tracking methods, such as MDNet and DCFST, showed the top performance in every year [15]. Not only that, the use of discriminative

correlation filters for localization and CNNs for feature extraction showed the top performance in recent year. In future, it is expected that CNN-based tracking methods will gain more popularity for tracking human tissues in thermal images because it has advantage to handle well blurring and low resolution as well as extremly deformation. Additionally novel methods of combining CNN and other conventional methods like DCF would be a direction of future research to improve the tracking performance further.

9 Conclusion

In general, thermal images show lower resolution, blurred edges and less texture over visible images. These features make it harder to find of human tissues and thus require a different approach compared to visible images. In the above sections, many methods from template updating as one of the classical methods to recent machine learning have been overviewed and evaluated for detecting vital signs in thermal images. Although machine learning methods showed better performance over classical methods in terms of speed, accuracy and robustness, they show pros and cons. Recently, many CNN-based tracking methods have been showing the top performance in every year in VOT contest. In future, CNN-based trackers are gaining popularity due to advantages of being able to handle blurring edges, low resolution and deformation better as well as to run in real time with recent powerful GPU technology, which are strongly required to track human tissues in thermal images for a long term. Furthermore, use of visual and thermal domain together for tracking can increase the performance significantly as the multi-modal data from RGB and thermal images will increase the amount of information gathers. Therefore, for more advanced tracking, the development of a cross-modality learning with the aid of CNN based deep learning should be studied deeper in near future [12].

References

[1] Iain Matthews, Takahiro Ishikawa, and Simon Baker, "The Template Update Problem." IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 26, no. 6, June 2004, pp. 810–815.

[2] Gade, Rikke, and Thomas B. Moeslund. "Thermal Cameras and Applications: A Survey." Machine Vision and Applications, vol. 25, no. 1, 9 Nov. 2013, pp. 245–262.

[3] Cootes, T.F, and C.J Taylor. "Active Shape Models - 'Smart Snakes'", BMVC, 1992, pp.266-275.

[4] Kalal, Zdenek, et al. "Forward-Backward Error: Automatic Detection of Tracking Failures." 2010 20th International Conference on Pattern Recognition, Aug. 2010, 10.1109/icpr.2010.675.

[5] Yan Zhou, Yan Zhou, Panagiotis Tsiamyrtzis, Peggy Lindner, Ilya Timofeyev, and Ioannis Pavlidis, "Spatiotemporal Smoothing as a Basis for Fa-cial Tissue Tracking in Thermal Imaging." IEEE Transactions on Biomedical Engineering, vol. 60, no. 5, May 2013, pp. 1280–1289.

[6] Allan D. Jepson David J. Fleet and Thomas F. El-Maraghi, "Robust Online Appearance Models for Visual Tracking." IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 25, no. 10, Oct. 2003, pp. 1296–1311.

[7] Marcin Kopaczka, Kemal Acar and Dorit Merhof, "Robust Facial Landmark Detection and Face Tracking in Thermal Infrared Images Using Active Appearance Models." Proceedings of the 11th Joint Conference on Computer Vi- sion, Imaging and Computer Graphics Theory and Applications, 2016, pp.150-158.

[8] Zdenek Kalal, Krystian Mikolajczyk, and Jiri Matas, "Tracking-Learning-Detection." IEEE Transactions on Pattern Analysis and Ma- chine Intelligence, vol. 34, no. 7, July 2012, pp. 1409–1422

[9] João F. Henriques, Rui Caseiro, Pedro Martins, and Jorge Batista, "High-Speed Tracking with Kernelized Correlation Filters", IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, 5 Nov. 2014.

[10] Milan Stojanovic, Natasa Vlahovic, Milos Stankovic and Srdan Stankovic, "Object Tracking in Thermal Imaging Using Kernelized Correlation Filters." 2018 17th International Symposium INFOTEH-JAHORINA (INFOTEH), Mar. 2018.

[11] Thi Zin, Hideya Takahashi, Takashi Toriu and Hiromitsu Hama, "Fusion of Infrared and Visible Images for Robust Person Detection", Image Fusion, Osamu Ukimura (Ed.),, 2011.

[12] Rother, Vladimir Kolmogorov and Andrew Blake, "GrabCut -Interactive Foreground Extraction Using Iter-ated Graph Cuts", 2004.

[13] Nam Hyeonseob, and Bohyung Han. "Learning Multi-Domain Convolutional Neural Networks for Visual Tracking". The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp 4293-4302.

[14] Zhu Gao, PoriKli, Fatih and Li, Hongdong. "Tracking Randomly Moving Objects on Edge Box Proposals". ArXiv, 29 Nov. 2015.

[15] Zitnick, C. and Dollár, P."Edge Boxes: Locating Object Proposals from Edges". ECCV, pp.391–405.2014.

[16] Anon, 2016. "The Thermal Infrared Visual Object Tracking VOT-TIR2016 Challenge Results".

[17] "The Visual Object Tracking VOT2015 challenge results". ICCV, 2015

[18] "The Seventh Visual Object Tracking VOT2019 Challenge Results". ICCV, 2019.